Lib.

Proceedings
NINTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

October 6-8, 1971

Allerton House, Monticello, Illinois
Sponsored by the
Department of Electrical Engineering and the
Coordinated Science Laboratory of the
University of Illinois at Urbana-Champaign

Allerton Conference on Circuit and
System Theory

9th Annual, 1971 [redacted]
[redacted]

[redacted]
LIBRARY

PROCEEDINGS

NINTH ANNUAL ALLERTON CONFERENCE ON CIRCUIT AND SYSTEM THEORY

J. B. Cruz, Jr.
A. H. Haddad

Conference Co-Chairmen

Conference held

October 6, 7, 8, 1971
Allerton House
Monticello, Illinois

Sponsored by the

DEPARTMENT OF ELECTRICAL ENGINEERING
and the
COORDINATED SCIENCE LABORATORY
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

[redacted]

FOREWORD

The Ninth Annual Allerton Conference on Circuit and System Theory was held at Allerton House on October 6-8, 1971. The Conference was co-sponsored by the Department of Electrical Engineering and the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign. These PROCEEDINGS contain the papers presented at the Conference.

In addition to the large number of submitted papers, there were a number of organized sessions which provided a nucleus for the Conference. These sessions were Estimation and Filtering, organized by A. H. Haddad; Estimation and Detection, organized by P. A. Frost; New Directions in Control System Theory and Applications, organized by M. K. Sain; Advances in Communications, organized by J. B. Thomas; Transportation Systems, organized by D. F. Wilkie; Modern Stability Theory and Its Applications to Power Systems, organized by P. V. Kokotović; Digital Filters, organized by J. K. Aggarwal; Environmental Remote Sensing, organized by R. Mittra; Parallel Processing Computers, organized by D. L. Slotnick. We are grateful to the organizers for their time and effort on behalf of the Conference. We wish to thank the authors in these sessions for their prompt response and excellent contribution to the Conference. We also acknowledge the participation of the IEEE Control Systems Society in the sessions organized by M. K. Sain and D. F. Wilkie. We are most grateful to the authors and the session chairmen for their effort and contribution to the success of the Allerton Conference.

We are indebted to Dean John E. Gibson for a very stimulating Keynote Address and to Professor B. R. Myers for his delightful introduction of the speaker for the ninth straight year. We thank Professor Jacob Wolfowitz for his interesting talk at the joint meeting with the Information Theory Group.

The support and encouragement of Professor C. L. Coates, Director of the Coordinated Science Laboratory, and Professor E. C. Jordan, Head of the Department of Electrical Engineering, are acknowledged and appreciated. We thank our colleagues for their generous assistance in organizing the Conference: C. L. Chen, R. T. Chien, P. V. Kokotović, W. Mayeda, G. Metze, W. R. Perkins, F. P. Preparata, T. N. Trick and N. Wax.

We are grateful to Mr. Harold Lawler for his handling of the business aspects of the Conference. We thank Sherry Kellenbach, Rose Lane, and Rosa Townsend for their help in the preparations of the Conference. Thanks are also due Captain R. Reinman for coordinating the projectionists and the visual aid equipment. Finally, special thanks are due to Mrs. Dee Keel for her expert and skillful coordination of the Conference arrangements.

J. B. Cruz, Jr.
A. H. Haddad

TABLE OF CONTENTS

Foreword iii

I-A: ESTIMATION AND FILTERING

Chairman: Y. C. Ho, Harvard University
Organizer: A. H. Haddad, University of Illinois at Urbana

ESTIMATION IN IDENTIFICATION THEORY 1
W. L. Root, University of Michigan, Ann Arbor

OPTIMAL AND SUBOPTIMAL RECURSIVE NONLINEAR FILTERING . . . 11
H. W. Sorenson, University of California at San Diego

SYNTHESIS OF SUB-OPTIMAL FIXED LAG SMOOTHERS 24
J. S. Meditch, University of California, Irvine

PARAMETER ESTIMATION IN PARTIALLY SPECIFIED STOCHASTIC SYSTEMS 32
R. L. Kashyap, Purdue University

I-B: OPTIMAL CONTROL

Chairman: John Zaborszky, Washington University, St. Louis, Mo.

PARAMETER DISCONTINUITIES IN OPTIMALLY CONTROLLED STOCHASTIC DISCRETE SYSTEMS 41
P. N. Nikiforuk, M. M. Gupta, and L. Wagner, University of Saskatchewan, Canada

GENERALIZED STOCHASTIC OPTIMIZATION WITH APPLICATIONS TO STOCHASTIC CONTROL AND OPTIMAL DESIGN OF RANDOM SYSTEMS 49
N. U. Ahmed, University of Ottawa, Canada

OPTIMAL CONTROL OF NON-UNIFORMLY SAMPLED DISCRETE SYSTEMS BY NONLINEAR PROGRAMMING 58
L. R. Nardizzi and B. K. Bhagavan, Southern Methodist University

OPTIMAL LINEAR REGULATORS WITH PRESCRIBED STEADY STATES. . 69
Milan S. Ćalović, University of Illinois at Urbana

A GRADIENT METHOD FOR THE OPTIMAL CONTROL PROBLEM OF BOLZA 82
S. J. Wang, GTE Laboratories, Inc., Waltham, Mass., and E. C. Tacker, Louisiana State University

NECESSARY CONDITIONS OF OPTIMALITY-A UNIFIED APPROACH. . . 84
V. P. Bhatkar and V. S. Rajamani, Indian Institute of Technology, New Delhi, India

I-C: COMPUTER ORGANIZATION

Chairman: S. S. Yau, Northwestern University

A STUDY OF DEMAND SCHEDULING IN A PARALLEL PROCESSING SYSTEM 93
Y. E. Chen and V. Y. Shen, Purdue University

A PIPELINE GRAPH FOR PARALLEL, ASYNCHRONOUS PROCESSING . . 102
R. Ribak, Worcester Polytechnic Institute, Worcester Mass.

THE WASTED SPACE-TIME INTEGRAL AND PAGE SIZE IN THE DESIGN OF VIRTUAL MEMORY COMPUTER SYSTEMS 104
E. Gelenbe and P. Tiberio, University of Michigan, Ann Arbor

TEST GENERATION USING A NETWORK FAULT EQUATION 114
J. W. Gault, North Carolina State University

CIRCUIT STRUCTURE AND SWITCHING FUNCTION VERIFICATION. . . 122
Min-Wen Du and C. D. Weiss, Johns Hopkins University

I-D: ACTIVE AND NONLINEAR CIRCUITS

Chairman: N. DeClaris, University of Maryland

A METHOD FOR QUICK DETERMINATION OF THE PERIODIC STEADY-STATE IN NONLINEAR NETWORKS 131
S. W. Director, A. J. Brodersen and D. A. Wayne, University of Florida, Gainesville

ABSOLUTE STABILITY OF NETWORKS CONTAINING MULTIPLE NONLINEAR ACTIVE ELEMENTS 140
C. Ho and Y. Jan, Rice University

MULTIPORT VOLTAGE GAIN FUNCTIONS: SYNTHESIS WITH ACTIVE RC MULTIPORT NETWORKS 142
T. A. Bickart, Syracuse University and D. W. Melvin, University of New Hampshire

HIGHER ORDER LINEAR CIRCUIT MODELING 150
J. S. Bedford and H. J. Neumann, The Pennsylvania State University

EXPERIMENTAL VERIFICATION OF SIGN REVERSAL OF TEMPERATURE COEFFICIENT 157
N. K. Bose, University of Pittsburgh and S. M. Kotob, American University of Beirut, Lebanon

I-E: SYSTEM ORDER REDUCTION

Chairman: W. A. Brown, Monash University, Australia (Visiting Professor at University of Waterloo, Canada)

ROUTH APPROXIMATION METHOD FOR HIGH-ORDER LINEAR SYSTEMS 160
M. F. Hutton, The Singer Company, Little Falls, N. J.

OPTIMUM APPROXIMATION OF HIGH-ORDER SYSTEMS BY LOW-ORDER MODELS USING RECENT GRADIENT METHODS 170
J. W. Bandler, N. D. Markettos and N. K. Sinha, McMaster University, Canada

ON ASYMPTOTIC SYSTEM THEORY: AN APPROACH TO SOME LARGE SCALE SYSTEMS 180
Celso de Renna e Souza and R. Saeks, COPPE/UFRJ, Rio de Janeiro, Brazil

ASYMPTOTIC EXPANSIONS OF SINGULARLY PERTURBED QUASI-LINEAR OPTIMAL SYSTEMS 182
P. Sannuti, Rutgers University

AN ALGORITHM FOR EQUIVALENT REDUCED STATE DESCRIPTIONS FOR LUMPED LINEAR TIME INVARIANT SYSTEMS 192
K. S. Bajaj, Michigan State University

II-A: ESTIMATION, DETECTION, AND FILTERING

Chairman and Organizer: P. A. Frost, Bell Telephone Laboratories, Whippany, N. J.

A NEW FILTER FOR OPTIMAL TRACKING IN DENSE MULTITARGET ENVIRONMENTS 201
R. A. Singer and R. G. Sea, Hughes Aircraft Company, Fullerton, California

SOME NEW RESULTS IN NON-GAUSSIAN DISCRETE-TIME ESTIMATION THEORY: RECURSIVE POLYNOMIAL ESTIMATORS 212
B. D. Jensen, Bell Telephone Laboratories, Whippany, N. J.

ESTIMATION FOR CYCLOSTATIONARY RANDOM PROCESSES 222
L. E. Franks and W. A. Gardner, University of Massachusetts

SOME SINGULAR ESTIMATION AND DETECTION PROBLEMS FOR CONDITIONALLY INDEPENDENT INCREMENT PROCESSES 232
P. A. Frost, Bell Telephone Laboratories, Whippany, N. J.

NONRECURSIVE SMOOTHING FOR GAUSSIAN MODULATED POINT PROCESSES 242
D. L. Snyder, Washington University School of Medicine, St. Louis, Missouri

JOINT DETECTION, ESTIMATION AND SYSTEM IDENTIFICATION 246
D. G. Lainiotis, University of Texas, Austin

II-B: NEW DIRECTIONS IN CONTROL SYSTEMS THEORY AND APPLICATIONS

Chairman and Organizer: M. K. Sain, University of Notre Dame

MATHEMATICAL MODELS FOR WARM AIR HEATING SYSTEMS 258
A. H. Eltimsahy, University of Toledo and
L. F. Kazda, University of Michigan, Ann Arbor

APPLICATION OF MODERN CONTROL THEORY TO GUIDANCE LAW DESIGN 272
D. M. Salmon, Systems Control, Inc., Palo Alto, California

NEW TRENDS IN THE CONTROL OF HEREDITARY SYSTEMS 282
H. N. Koivo, University of Toronto, Canada

RESOLUTION SPACE - A FUNCTION ANALYTIC SETTING FOR CONTROL THEORY 290
R. Saeks, University of Notre Dame

II-C: AUTOMATA AND SEQUENTIAL MACHINES

Chairman: D. E. Muller, University of Illinois at Urbana

GENERATION OF REGULAR EXPRESSIONS FOR AUTOMATA BY THE INTEGRAL OF REGULAR EXPRESSIONS 299
L. W. Smith and S. S. Yau, Northwestern University

ON AUTOMATA WITH FINITE REFERENCED PATTERNS 309
E. J. Krall, University of Dayton and R. T. Yeh, University of Texas

CONSTRUCTION OF A SEQUENTIAL MACHINE MAPPING A GIVEN INPUT REGULAR EXPRESSION ONTO A GIVEN OUTPUT REGULAR EXPRESSION 318
R. O. Lind and S. S. Yau, Northwestern University

A NOTE ON THE AUTONOMOUS BEHAVIOR OF FINITE STATE MACHINES 329
A. Miczo, IBM Corp., Owego, N. Y. and L. D. Rudolph, Syracuse University

REALIZATION OF THRESHOLD SEQUENTIAL MACHINES BASED ON A MODIFICATION OF THE THEORY OF COMPLETE MONOTONICITY . . 330
S. G. Burgiss, Terminal Communications, Inc., Raleigh, N. C. and John Staudhammer, North Carolina State University

ELIMINATION OF SPURIOUS OUTPUT PULSES FOR MOORE TYPE SEQUENTIAL MACHINES 338
J. G. Bredeson and P. T. Hulina, Pennsylvania State University

TRIGGER MACHINES: CANONICAL FORMS AND OTHER PROPERTIES . 347
H. T. Hao, IBM Advanced Systems Dev. Div., Yorktown Heights, N. Y. and M. M. Newborn, Columbia University

ALGORITHMS FOR CONSTRUCTING TRANSITION TABLES OF MINIMUM-STATE COMPUTING SEQUENTIAL MACHINES 357
S. C. Lee, University of Houston

II-D: OPTIMIZATION METHODS

Chairman: K. Steiglitz, Princeton University

A HEURISTIC APPROACH TO MULTIMODAL FUNCTION MINIMIZATION - THE GLOBAL ASPECT 369
J. Opačić, University of Maryland

EVOLUTIONARY LEARNING OF NEIGHBORHOODS FOR HEURISTIC PROGRAMS AND APPLICATION TO A SEQUENCING PROBLEM 377
W. H. Kohler and K. Steiglitz, Princeton University

ON RANDOM SEARCH STRATEGIES FOR PARAMETER OPTIMIZATION. . 390
M. J. Wozny, Oakland University, Rochester, Michigan

A SECOND ORDER GENERALIZED REDUCED GRADIENT TECHNIQUE . . 397
L. Meier, Systems Control, Inc., Palo Alto, California

ON CONDITIONS FOR OPTIMALITY IN LEAST pTH APPROXIMATION WITH $p \to \infty$ 404
J. W. Bandler and C. Charalambous, McMaster University, Canada

TWO-PARAMETER FUNCTIONALS FOR PROGRAMMING IN NORMED SPACES 414
R. H. Smith, Jr. and V. D. Vandelinde, The Johns Hopkins University

AN ITERATIVE VECTOR SPACE METHOD FOR EXPONENTIAL APPROXIMATION 424
R. K. Harman and F. W. Fairman, Queen's University, Canada

II-E: SYSTEMS MODELING AND SIMULATION

Chairman: D. D. Šiljak, University of Santa Clara

TEMPORAL PROPERTIES OF ENGINEERING SYSTEMS 434
R. M. DeSantis, Ecole Polytechnique, Montreal, Canada and W. A. Porter, University of Michigan, Ann Arbor

URBAN HEALTH CARE SIMULATION: AN ECOLOGICAL APPROACH TO SYSTEM ANALYSIS AND EVALUATION 444
J. Milsum, McGill University, Canada, R. Strohmaier, D. Uyeno, I. Vertinsky and H. Will, University of British Columbia

ON NEURON MODELING . . . .
N. DeClaris, University of Maryland

A COMPUTER APPROACH FOR ANALYSIS OF NONLINEAR CIRCUITS. . 455
D. A. R. Zein and S. L. Hakimi, Northwestern University

HIGH ORDER A-STABLE COMPOSITE MULTISTEP METHODS FOR NUMERICAL INTEGRATION OF STIFF DIFFERENTIAL EQUATIONS . . 465
T. A. Bickart and Donald A. Burgess, Syracuse University and H. M. Sloate, General Electric Electronics Laboratory, Syracuse, N. Y.

A DIGITAL SIMULATION METHOD FOR INTERCONNECTED CONTINUOUS SYSTEMS 474
R. A. DiPerna, Philips Industries, Croydon, England

III-A: ADVANCES IN COMMUNICATIONS

Chairman and Organizer: J. B. Thomas, Princeton University

RECENT ADVANCES IN DATA COMMUNICATION 476
H. O. Burton, Bell Telephone Laboratories, Holmdel, N. J.

A SURVEY OF DATA COMPRESSION 477
L. D. Davisson, University of Southern California, Los Angeles

A NEW LOOK AT THE MATHEMATICAL FOUNDATIONS OF COMMUNICATION THEORY 485
T. T. Kadota, Bell Telephone Laboratories, Murray Hill, N. J.

ACHIEVEMENTS IN IMAGE PROCESSING 497
O. J. Tretiak, Massachusetts Institute of Technology

NOISELESS CODING FOR SIGNAL DESIGN 507
P. A. Franaszek, IBM Corp., Yorktown Heights, N. Y.

III-B: TRANSPORTATION SYSTEMS I

Chairman and Organizer: D. F. Wilkie, Ford Motor Co.

COMPUTER REPRESENTATION AND MANIPULATION OF TRANSPORTATION NETWORKS 514
V. J. Marma, Transportation Systems Center, Cambridge, Mass.

OPTIMIZATION OF PROGRESSION TIMINGS ON A UNIFORM GRID STREET NETWORK 525
J. B. Kreer, Michigan State University

RESERVATION SYSTEMS FOR SYNCHRONOUS TYPE GUIDEWAY AUTOMATION 531
L. T. Klauder, Jr., Louis T. Klauder and Associates, Philadelphia, Pa.

A SYSTEM ANALYSIS OF UNIVERSITY TRANSPORT NEEDS 538
J. E. Gibson and A. W. Turski, Oakland University, Rochester, Michigan

NETWORK EFFECTS ON THE INTERFACE PROBLEM FOR AN AUTOMATED HIGHWAY SYSTEM 547
R. G. Stefanek, Ford Motor Company

A STUDY OF THE CONTROL OF MAGNETICALLY LEVITATED VEHICLES 559
D. F. Wilkie, Ford Motor Company

III-C: CONTROL SYSTEMS DESIGN

Chairman: Captain M. D. Ciletti, F. J. Seiler
Research Laboratory, USAF Academy,
Colorado

A MAXIMIZED LIAPUOV FUNCTION FOR POWER SYSTEM STABILITY STUDY 576
Y. N. Yu and A. A. Metwally, University of British Columbia

A COMPUTER-AIDED REGULATOR DESIGN 585
J. S. Karmarkar, Ames Research Center (NASA), Moffett Field, California and D. D. Šiljak, University of Santa Clara

SUBOPTIMAL COMPENSATOR DESIGN FOR THE LINEAR REGULATOR VIA SECOND-ORDER SENSITIVITY 595
W. A. Brown, Monash University, Australia (on leave at the University of Waterloo, Canada)

AN IMPROVED DESIGN TECHNIQUE FOR PARAMETER ADAPTIVE CONTROL SYSTEMS 604
J. W. Gilbart, Goddard Space Flight Center, Greenbelt, Md. and R. V. Monopoli and E. Lizewski, University of Massachusetts

DESIGN OF A DIGITAL ADAPTIVE CONTROL SYSTEM FOR REENTRY VEHICLES 613
J. L. Picon-Jimenez and R. C. Montgomery, Langley Research Center, Hampton, Virginia and L. L. Grigsby, VPISU, Blacksburg, Virginia

ON PULSED LASER SPOT HOMING MISSILES 624
E. T. Evers-Euterneck, U. S. Army Missile Command, Redstone Arsenal, Alabama

III-D: NETWORK THEORY

Chairman: B. J. Leon, Purdue University

FUNDAMENTAL CUT-SET CONTOURS AND REALIZATION OF CUT-SET MATRICES 634
M. Lal and N. C. Jain, University of Roorkee, India

SOME DUALITY RELATIONSHIPS IN NETWORKS 644
M. I. Dessouky, University of Illinois at Urbana and E. J. Dunne, Wright-Patterson Air Force Base, Dayton, Ohio

STRATEGIES FOR TEARING MULTITERMINAL NETWORKS 652
W. W. Happ, Sacramento State College and L. P. McNamee, University of California, Los Angeles

GENERALIZED INVERSES, DUALS, DUAL ADJOINTS AND APPLICATIONS 661
M. N. S. Swamy, C. Bhushan and B. B. Bhattacharyya, Sir George Williams University, Canada

STATE-SPACE INTERPRETATION OF CLASSICAL RESULTS IN NETWORK THEORY 663
M. Lal and H. Singh, University of Roorkee, India

OPTIMUM SENSITIVITY DESIGN FOR A CLASS OF ACTIVE FILTERS 673
W. S. Wagner, University of Cincinnati

MINIMIZATION OF THE QUALITY FACTOR SENSITIVITY 675
C. S. Phan and H. K. Kim, University of Manitoba, Canada

III-E: CODING THEORY

Chairman: L. D. Rudolph, Syracuse University

A NEW LOOK AT PERIODIC CODES 677
D. J. Costello, Jr., Illinois Institute of Technology

ON THE MINIMUM DISTANCE STRUCTURE OF CYCLIC CODES OF COMPOSITE LENGTH 685
C. R. P. Hartmann, Syracuse University and K. K. Tzeng, Lehigh University

A NOTE ON THE MINIMUM DISTANCE STRUCTURE OF CYCLIC CODES 687
C. R. P. Hartmann, Syracuse University

SOME NOTES ON ITERATIVE DECODING 689
K. K. Tzeng, Lehigh University, C. R. P. Hartmann, Syracuse University and R. T. Chien, University of Illinois at Urbana

DETERMINING CODE CAPABILITIES FOR NONRANDOM ERRORS . . . 696
D. A. Anderson, Bell Telephone Laboratories, Naperville, Illinois and G. Metze, University of Illinois at Urbana

SYSTEMATIC NON-SEPARATE ARITHMETIC GROUP CODES 709
O. N. Garcia and R. T. Greene, University of South Florida, Tampa

IV-A: MODERN STABILITY THEORY AND ITS APPLICATIONS TO POWER SYSTEMS

Chairman and Organizer: P. V. Kokotović, University of Illinois at Urbana

POWER SYSTEM CONTROL PRACTICE 719
N. Cohn, Leeds & Northrup Co., North Wales, Pa.

HIGH VOLTAGE DIRECT CURRENT TRANSMISSION (HVDC) AND POWER SYSTEM STABILITY 730
V. Madzarevic, E. G. Pereboom, and J. Sabath, Southern California Edison Co., Los Angeles

ON LARGE-SCALE SYSTEM STABILITY 731
D. D. Šiljak, University of Santa Clara

THE APPLICATION OF FREQUENCY DOMAIN STABILITY CRITERIA TO THE CONSTRUCTION OF LYAPUNOV FUNCTIONS FOR POWER SYSTEMS 742
J. L. Willems, University of Gent, Belgium

CRITICAL SURVEY OF TRANSIENT STABILITY STUDIES OF MULTIMACHINE POWER SYSTEMS BY LIAPUNOV'S DIRECT METHOD . 751
M. Ribbens-Pavella, University of Liège, Belgium

POWER SYSTEM STATE REGULATOR DESIGN WITH DOMINANT EIGENVALUE SHIFT 768
Y. N. Yu and H. Moussa, University of British Columbia

IV-B: DIGITAL FILTERS

Chairman and Organizer: J. K. Aggarwal, University of Texas, Austin

DESIGNING SHORT-WORD RECURSIVE DIGITAL FILTERS 778
K. Steiglitz, Princeton University

ON OPTIMUM NONRECURSIVE DIGITAL FILTERS 789
E. M. Hofstetter, A. V. Oppenheim and J. Siegel, Massachusetts Institute of Technology

DIGITAL FILTERING WITH VARIOUS INTEGRATION ROUTINES . . . 799
B. J. Leon, Purdue University and T-C. Huang, University of Detroit

SOME DESIGN CONSIDERATIONS IN THE IMPLEMENTATION OF DIGITAL FILTERS 809
J. F. Kaiser, Bell Telephone Laboratories, Murray Hill, N. J.

ON CANONIC REALIZATIONS OF DIGITAL FILTERS 810
S. K. Mitra and R. Sherwood, University of California, Davis

IV-C: ECONOMIC SYSTEMS

Chairman: L. F. Kazda, University of Michigan, Ann Arbor

DYNAMICS OF LESS DEVELOPED COUNTRIES 827
K. Chen and H. E. Hoelscher, University of Pittsburgh

EXPERIMENTS WITH A DYNAMIC MODEL OF A LESS DEVELOPED COUNTRY 837
K. Oka, University of Pittsburgh

A DYNAMIC INPUT-OUTPUT MODEL FOR AN ECONOMIC SYSTEM . . . 845
W. G. Vogt, M. H. Mickle and H. Aldermeshian, University of Pittsburgh

SUFFICIENT CONDITIONS FOR ALMOST CERTAIN STABILITY AND ASYMPTOTIC STABILITY OF A CLASS OF REGENERATIVE POPULATION SYSTEMS 855
W. D. O'Neill, University of Illinois at Chicago

AN APPLICATION OF D-DECOMPOSITION TO THE STABILITY ANALYSIS OF A POPULATION MODEL 862
N. Jermihov, University of Illinois at Chicago

AN INTERACTIVE FACILITY FOR EXAMINING DEVELOPMENT POLICIES OF LOCAL GOVERNMENT 864
E. J. Cristiani, R. E. Goldman, Santa Clara County (Calif.) Planning Dept. and R. J. Evey and P. E. Mantey, IBM Research Division, San Jose, California

IV-D: LINEAR SYSTEM THEORY

Chairman: K. Yao, University of California, Los Angeles

DECOUPLING IN LINEAR TIME-VARIABLE SYSTEMS USING OUTPUT FEEDBACK 866
S. N. Singh and W. J. Rugh, The Johns Hopkins University

POLE ASSIGNMENT BY MODE FEEDBACK 876
H. Hemami, The Ohio State University

A NEW ALGORITHM FOR THE IRREDUCIBLE REALIZATION OF A RATIONAL MATRIX 884
C. G. Giorgi and A. Isidori, Universita di Roma, Italy

ON BANDPASS FUNCTIONS FROM LOWPASS POLES 894
H. Gordon, Technical Consultant, Sarasota, Florida and G. W. Zobrist, University of South Florida, Tampa

A CRITERION TO DETERMINE IF TWO MULTIVARIABLE POLYNOMIALS ARE RELATIVELY PRIME 901
N. K. Bose, University of Pittsburgh

AN OPTIMAL ITERATIVE METHOD FOR THE COMPUTATION OF MATRIX RANK 906
N. N. Gupta, Wright State University, Dayton, Ohio

IV-E: DIFFERENTIAL GAMES

Chairman: J. E. Heller, Bell Telephone Laboratories, Murray Hill, N. J.

A NOTE ON THE GENERALIZED HAMILTON-JACOBI EQUATION FOR DIFFERENTIAL GAMES WITH INFORMATION TIME LAG 911
M. D. Ciletti, F. J. Seiler Research Laboratory, USAF Academy, Colorado

ON RENDEZVOUS AND EVASION OF LINEAR SYSTEMS WITH MULTIPLE CONSTRAINTS ON CONTROLLERS 921
D. H. Chyung, University of Iowa

OPTIMUM CONTROL IN A DISTRIBUTED SYSTEM UNDER CONFLICT. . 924
M. C. Y. Kuo, Newark College of Engineering

A TWO PERSON ZERO SUM STOCHASTIC GAME THAT ADMITS A MAXIMIN POLICY DEFINED BY A GUASSIAN MEASURE 932
T. Basar and M. Mintz, Yale University

LINEAR DIFFERENCE GAMES WITH DELAYS 942
J. L. Hanson, Air Force Institute of Technology, Wright-Patterson AFB, Ohio

V-A: ENVIRONMENTAL REMOTE SENSING

Chairman: K. S. Fu, Purdue University
Organizer: R. Mittra, University of Illinois at Urbana

PATTERN RECOGNITION IN REMOTE SENSING 951
K. S. Fu, Purdue University

IMAGING RADAR TECHNIQUES FOR REMOTE SENSING APPLICATIONS. 975
J. S. Zelenka, University of Michigan

TOOLS FOR REMOTE SENSING EXPERIMENTS 987
G. Nagy, IBM Corp., Yorktown Heights, N. Y.

SOME ANALYTICAL AND NUMERICAL TECHNIQUES FOR REMOTE SENSING 997
R. Mittra, D. H. Schaubert and M. Mostafavi, University of Illinois at Urbana

V-B: PARALLEL PROCESSING COMPUTERS

Chairman and Organizer: D. L. Slotnick, University of Illinois at Urbana

BACKGROUND AND HISTORY OF PARALLEL PROCESSING 1007
D. L. Slotnick, University of Illinois at Urbana

ILLIAC IV HARDWARE STRUCTURE 1010
S. A. Denenberg, University of Illinois at Urbana

ILLIAC IV SOFTWARE: AN OVERVIEW 1020
J. M. Randal, University of Illinois at Urbana

ILLIAC IV APPLICATIONS 1030
A. H. Sameh, University of Illinois at Urbana

ILLIAC IV APPLICATIONS AND CURRENT STATUS 1039
D. E. McIntyre, University of Illinois at Urbana

ILLIAC IV AND THE ARPA NETWORK 1045
W. J. Bouknight, University of Illinois at Urbana

V-C: SIGNAL PROCESSING IN COMMUNICATIONS

Chairman: L. E. Franks, University of Massachusetts

ON OPTIMAL PAM MODULATION AND DEMODULATION 1055
K. Murano and J. B. Thomas, Princeton University

THRESHOLD PERFORMANCE OF A PHASE-LOCKED LOOP WITH PHASE FEEDBACK 1067
J. Klapper, Newark College of Engineering and G. M. Veiga, Western Electric Co., Newark, N. J.

A STUDY OF A BANDWIDTH-ADAPTIVE SPEECH PROCESSOR 1076
T. E. Eger, Quest Research Corp., McLean, Va.

ESTIMATION OF SPECTRAL MOMENTS 1085
J. N. Denenberg, Bell Telephone Laboratories, Naperville, Ill. and L. C. Peach, Illinois Institute of Technology

THE LAW OF LARGE NUMBERS FOR CYCLOSTATIONARY PULSE TRAINS 1094
R. Lugannani, Princeton University

ON NUMERICAL CALCULATIONS OF BOUNDS OF THE CAPACITY OF A TIME-DISCRETE GAUSSIAN CHANNEL WITH TIMING JITTER . . . 1104
K. Yao, University of California, Los Angeles

ON COMPUTING THE PHASE SPECTRUM FROM THE DISCRETE FOURIER TRANSFORM OF A SAMPLED SIGNAL 1112
H. Schwarzlander, Syracuse University

V-D: TRANSPORTATION SYSTEMS II

Chairman: D. M. Salmon, Systems Control, Palo Alto, Calif.

REGULATION PERFORMANCE OF A CAR-FOLLOWING ALGORITHM IN THE PRESENCE OF NOISE 1114
V. D. Vandelinde, The Johns Hopkins University

AN APPROACH TO MERGING CONTROL IN GROUND TRANSPORTATION SYSTEM 1124
H. E. Lee, Westinghouse Electric Corp., Annapolis, Maryland and D. W. C. Shen, University of Pennsylvania

THE OPTIMAL CONTROL OF MERGING AIRCRAFT-DERIVATION OF THE HYBRID AIR TRAFFIC CONTROLLER 1130
J. G. Schatz, Bell Aerospace, Buffalo, N. Y. and T. E. Stern, Columbia University

A STOCHASTIC NETWORK MODEL FOR AIR CARGO TERMINALS . . . 1140
C. Lee, Litton Data Systems Div., Woodland Hills, Calif. and L. P. McNamee, University of California, Los Angeles

ACTIVITY NETWORKS TO MODEL TRANSPORTATION SYSTEMS SUBJECT TO FACILITY CONSTRAINTS 1151
H. A. Porte, U. S. Army Corps. of Engineers and W. W. Happ, University of Illinois at Urbana

V-E: DISTRIBUTED NETWORKS AND SYSTEMS

Chairman: B. A. Shenoi, University of Minnesota

STABILITY OF DISTRIBUTED NETWORKS WITH NONLINEAR ELEMENTS 1161
P. J. Schwyn, IBM Corp., Endicott, N. Y. and T. A. Bickart, Syracuse University

APPLICATION OF DISTRIBUTED NETWORK SYNTHESIS PROCEDURES TO THE SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS 1170
R. D. Hensley and J. J. Bourquin, University of Missouri-Rolla

ON A NEW CHARACTERIZATION OF TWO PORT DISTRIBUTED NETWORK IMMITTANCE MATRICES 1180
K. S. Bajaj and J. A. Resh, Michigan State University

PROPERTIES OF NETWORK PARAMETERS OF COUPLED NON-UNIFORM LC LINES 1191
K. Singhal and J. Vlach, University of Waterloo, Canada

DISCRETE DISTRIBUTED SYSTEMS 1193
G. L. Kusic, University of Pittsburgh and
V. Arandjelovic, "Boris Kidric" Institute of
Nuclear Sciences, Belgrade

A SENSOR LOCATION ALGORITHM FOR DISTRIBUTED-PARAMETER SYSTEMS 1203
D. J. Ewing, Jr., University of Toledo and
T. J. Higgins, University of Wisconsin

BOUNDARY FILTER FOR LINEAR DISTRIBUTED PARAMETER SYSTEMS 1213
S. R. Atre and S. S. Lamba, Indian Institute of
Technology, New Delhi, India

INDEX OF AUTHORS 1222

ESTIMATION IN IDENTIFICATION THEORY

W. L. ROOT
Computer, Information & Control Engineering
The University of Michigan, Ann Arbor, Michigan

ABSTRACT

Many problems in identification and measurement are conveniently formulated in a Hilbert space format, leading to infinite-dimensional statistical estimation problems which often pose convergence difficulties. Some of the usual estimation procedures are unsatisfactory even for finite-dimensional approximations. A simple linear estimator which is suitable for a class of such problems is presented.

INTRODUCTION

A Class of Identification Problems

Let a system represented by a transformation H have input x and output y, and suppose there is additive observation noise n Then the system is represented by the equation

$$y = H(x) + n \tag{1}$$

Suppose the system transformation H is unknown, although known to belong to some class $\mathcal{C}$. Then one can pose the identification problem: find $H \in \mathcal{C}$ from a knowledge of the input x and observation of the output y. This is an important and standard sort of problem with an extensive literature, but I refer specifically only to a very special part of it (1), (2), (3), (4), (5), (6), (7), because the point of view developed in these papers leads to the statistical estimation problem to be discussed here. Very briefly, the identification theory referred to treats the following situation: x and y belong to, say, certain function spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and H is a continuous mapping from $\mathcal{X}$ into $\mathcal{Y}$ (not necessarily linear, nor necessarily time-invariant, although these are important special cases that have been studied in detail). It is stipulated that, given $\epsilon > 0$, H is to be found to within ϵ in some norm imposed on a class $\mathcal{H}$ of system transformations containing $\mathcal{C}$. It is further stipulated that the computation yielding the approximation to H must depend on only a finite number of observed values (real numbers) related to the output. A class $\mathcal{C}$ of transformations H for which this is possible for any ϵ for the noise-free case is called a determinable class. One of the conditions a determinable class $\mathcal{C}$ must satisfy if the H are uniformly bounded is that it be conditionally compact. If repeated measurements are possible this is also sufficient, at least in a rather general case, see (7). Thus, even before considering the output noise, n, one is led to impose the restriction that $H \in \mathcal{C}$, $\mathcal{C}$ a subset of the continuous bounded mappings from $\mathcal{X}$ into $\mathcal{Y}$ which is compact in the topology given by the norm used to measure the error.

Now the class $\mathcal{H}$ of all continuous bounded mappings from $\mathcal{X}$ into $\mathcal{Y}$ is always a linear space with addition and scalar multiplication of functions defined in the usual way. It is immediate that any fixed x defines a linear mapping X from the linear span of $\mathcal{C} \subset \mathcal{H}$ into $\mathcal{Y}$ defined by

$$XH = H(x), \quad H \in \mathcal{C} \tag{2}$$

Thus, the problem of inverting the equation $y = H(x) = XH$ for H is always a linear one (unless a non-linear representation is deliberately introduced, see (7)), and the related statistical problem of estimating H in $y = H(x) + n = XH + n$ is also linear.

The Resulting Class of Estimation Problems

The condition that the identification must follow from a finite number of observations means that ultimately the output space $\mathcal{Y}$ is replaced by a finite-dimensional one, the dimension depending on the ϵ prescribed for the approximation. Thus we are led to consider the problem of estimating H in systems represented by

$$z = P_\epsilon y = P_\epsilon XH + P_\epsilon n \tag{3}$$

where the P_ϵ are projections with finite-dimensional range in $\mathcal{Y}$, where X is a linear transformation (which is always assumed bounded, but will usually not be bounded from below), and where $H \in \mathcal{C}$, a prescribed compact subset of $\mathcal{H}$.

For the remainder of this paper we take $\mathcal{Y}$ to be a separable Hilbert space, $\mathcal{X}$ to be a subset of a separable Hilbert space and $\mathcal{H}$ also to be a separable Hilbert space. This is considerably more restrictive than necessary (see (7)), but it is sufficient, for example, to encompass most problems where H is represented as a Volterra polynomial operator and $\|H\|$ is a generalized Hilbert-Schmidt norm of the operator (see (3), (6)). An example using such a representation is given in the last section.

The estimation problem thus posed presents no real difficulties and presumably can be solved satisfactorily in a number of ways. The estimators developed in this paper can be regarded as modifications of linear unbiased minimum variance (lumv) estimators. The mathematical theory of these new estimators is nearly trivial; the justification for discussing them is that they appear to be useful.

In the next Section there is a brief summary of facts about lumv estimators. They are set in the context of infinite-dimensional Hilbert spaces, but only the easy special cases are discussed. In the last Section the new estimators are discussed and an example is given.

STANDARD RESULTS ABOUT LUMV ESTIMATORS

Notation and Preliminaries

We consider the equation

$$y = Xh + n \tag{4}$$

where y is an element of a real separable Hilbert space $\mathcal{Y}$ (which may be finite dimensional), h is an element of a real separable Hilbert space $\mathcal{H}$ (which also may be finite dimensional), X is a bounded linear transformation from $\mathcal{H}$ into $\mathcal{Y}$, and n is a random variable taking values in $\mathcal{Y}$. It is assumed that n has mean zero, and that its covariance operator R exists (R is the unique operator, if it exists, satisfying $E<n,f><n,g> = <Rf,g>$ for all $f,g \in \mathcal{Y}$). R is always Hermitian and non-negative definite. We sometimes require that R be bounded from below. In the finite-dimensional case this simply means that R is strictly definite, but in the infinite-dimensional case it means more, of course, and there is a difficulty. If R is bounded from below it cannot have finite trace, which is a necessary and sufficient condition that R be the covariance operator of a countably additive probability measure on the Borel sets of $\mathcal{Y}$. The difficulty can be circumvented by supposing that n is a "basic" random variable defined in terms of a finitely additive measure (see (8)).

Before discussing lumv estimation it is convenient to note some basic relationships involving the adjoint of X and certain other operators formed from X, its adjoint and R. Let the null space of X be denoted by $\mathcal{N}(X)$ and the range by $\mathcal{R}(X)$. That is, $\mathcal{N}(X) = \{h: Xh = 0\}$, $\mathcal{R}(X) = \{y: y = Xh$ for some $h \in \mathcal{H}\}$. The adjoint of X, X^*, is the bounded linear transformation from $\mathcal{Y}$ to $\mathcal{H}$ defined by $< Xh,z > \; = \; < h,X^*z >$ for all $h \in \mathcal{H}$, $z \in \mathcal{Y}$. Here, $< Xh,z >$ is an inner product in $\mathcal{Y}$, whereas $< h,X^*z >$ is an inner product in $\mathcal{H}$ - there is no real ambiguity in using the same notation for inner products (or norms) in both spaces.

X^*X is a self-adjoint, non-negative bounded linear operator on $\mathcal{H}$ (i.e., a transformation from $\mathcal{H}$ into $\mathcal{H}$), and XX^* is a self-adjoint, non-negative bounded linear operator on $\mathcal{Y}$. By $(X^*X)^{-1}$, for example, we mean the linear operator defined only on the range of X^*X such that

$$X^*X(X^*X)^{-1}h = (X^*X)^{-1}X^*Xh = h$$

for all $h \in \mathcal{R}(X^*X)$. If $\mathcal{H}$ is infinite-dimensional $(X^*X)^{-1}$ may be unbounded, of course. $(XX^*)^{-1}$ is to be defined analogously. The following facts are standard, and can be proven easily:

(A) 1. $\mathcal{N}(X)$ and $\mathcal{N}(X^*)$ are closed linear subspaces of $\mathcal{H}$ and $\mathcal{Y}$, respectively.

2. $\mathcal{R}(X)$ and $\mathcal{R}(X^*)$ are linear subsets of $\mathcal{Y}$ and $\mathcal{H}$, respectively. They are not necessarily closed in general. They are closed when $\mathcal{Y}$ and $\mathcal{H}$ are finite-dimensional.

3. Let $\overline{\mathcal{R}(X)}$ be the closure of $\mathcal{R}(X)$, and $\oplus$ denote direct sum. Then $\mathcal{H} = \mathcal{N}(X) \oplus \overline{\mathcal{R}(X^*)}$, $\mathcal{Y} = \mathcal{N}(X^*) \oplus \overline{\mathcal{R}(X)}$.

4. $\mathcal{N}(X^*X) = \mathcal{N}(X)$, $\mathcal{N}(XX^*) = \mathcal{N}(X^*)$.

5. $\mathcal{R}(X^*X) \subset \mathcal{R}(X^*) \subset \overline{\mathcal{R}(X^*X)}$, $\mathcal{R}(XX^*) \subset \mathcal{R}(X) \subset \overline{\mathcal{R}(XX^*)}$.

6. If any one of the sets $\mathcal{R}(X)$, $\mathcal{R}(X^*)$, $\mathcal{R}(X^*X)$, $\mathcal{R}(XX^*)$ are closed, so are all of the others.

If X is a compact transformation, then X^*, X^*X and XX^* are compact. Since X^*X and XX^* are also self-adjoint and non-negative they have the discrete spectral decomposition given by the spectral theorem for such operators. We then have:

(B) Let $\{\phi_i\}$ be a full set of orthonormal eigenvectors of X^*X with corresponding non-zero eigenvalues λ_i, $X^*X\phi_i = \lambda_i\phi_i$, $\lambda_i > 0$. Then if $\eta_i = \lambda_i^{-\frac{1}{2}}X\phi_i$, it follows that $XX^*\eta_i = \lambda_i\eta_i$, $< \eta_i,\eta_j > \; = \delta_{ij}$, $\phi_i = \lambda_i^{-\frac{1}{2}}X^*\eta_i$. Also $\mathcal{N}(X^*X)$ is the orthogonal complement of the subspace spanned by the ϕ_i, and $\mathcal{N}(XX^*)$ is the orthogonal complement of the subspace spanned by the η_i.

Since the operator R is Hermitian and non-negative definite, it has a Hermitian, non-negative definite square root, $R^{\frac{1}{2}}$. By $R^{-\frac{1}{2}}$ we mean

$(R^{\frac{1}{2}})^{-1} = (R^{-1})^{\frac{1}{2}}$ If R is bounded and strictly positive definite the ranges of R^{-1} and $R^{-\frac{1}{2}}$ are $\mathcal{Y}$. Then $\mathcal{R}(X^*R^{-\frac{1}{2}}) = \mathcal{R}(X^*)$ so $\mathcal{R}(X^*R^{-\frac{1}{2}})$ is closed if $\mathcal{R}(X^*)$ is closed. We shall need to apply the facts listed in (A) and (B) with X replaced by $R^{-\frac{1}{2}}X$, so that X^*X is replaced by $X^*R^{-1}X$.

LUMV Estimates

An lumv estimator of h in Eq. (4) is a random variable $\hat{h}$ taking values in $\mathcal{H}$ and of the form $\hat{h} = Cy$, where C is a bounded linear mapping from $\mathcal{Y}$ into $\mathcal{H}$, such that $E\hat{h} = h$ (unbiased), and such that $E\|\hat{h} - h\|^2 = \min\limits_{\{\tilde{h}\}} E\|\tilde{h} - h\|^2$, where $\tilde{h}$ is a linear unbiased estimator (minimum variance). An lumv estimate may not exist: there may be no linear estimates that are unbiased, and in the infinite-dimensional case there may be linear unbiased estimates, but none with finite variance.

The following facts are well-known or are proved easily from known results.

(C) 1. If there is no supplementary information about the unknown h, there can be no unbiased estimator of h if $\mathcal{N}(X) \neq 0$. (This is obvious, and true in general.)

2. Let h_1 be the orthogonal projection of h into $\mathcal{N}(X)^{\perp}$. Suppose that the random variable n in $\mathcal{Y}$ has mean zero and a bounded covariance operator R on $\mathcal{Y}$ that is strictly positive definite. Suppose,

i) $\mathcal{R}(X^*) = \mathcal{R}(X^*R^{-\frac{1}{2}})$ is closed in $\mathcal{H}$,

ii) $(X^*R^{-1}X)^{-1}X^*R^{-1}$ is densely defined on $\mathcal{Y}$ and bounded, and hence has a bounded extension C defined on all of $\mathcal{Y}$,

iii) $\mathrm{Tr}[(X^*R^{-1}X)^{-1}]$ exists ($\mathrm{Tr}(\cdot)$ denotes trace). Then

$$\hat{h}_1 = Cy \qquad (5)$$

is an lumv estimator of h_1 with variance $E\|\hat{h}_1 - h_1\|^2 = \mathrm{Tr}[(X^*R^{-1}X)^{-1}]$.

3. If $\mathcal{R}(X^*)$ is finite-dimensional and R is strictly positive definite, all the conditions of 2. are satisfied and $\hat{h}_1$ as given by Eq. (5) is an lumv estimator of h_1.

Comments:

(a) The conditions of D(2) can be satisfied in the infinite-dimensional case, but it is necessary that X be bounded from below.

(b) $\mathcal{R}(X^*)$ is of finite dimension iff $\mathcal{R}(X)$ is of finite dimension (and the dimensions are the same). Hence if $\mathcal{Y}$ and $\mathcal{H}$ are arbitrary Hilbert spaces, but P and Q are, respectively, orthogonal projections in $\mathcal{Y}$ and $\mathcal{H}$

with finite-dimensional ranges, then the restricted estimation problems given by the following equations are all finite-dimensional, in the sense of (C)3. above.

$$y' = Py = PXh + Pn \tag{6}$$

$$y'' = XQh + n \tag{7}$$

$$y''' = Py'' = PXQh + Pn \tag{8}$$

Consider Eq. (7), for example. We have

$$\hat{h}_1 = (QX^* R^{-1} XQ)^{-1} QX^* R^{-1} y''$$

as lumv estimator of Qh_1, and

$$E\| h_1 - Qh_1\|^2 = \mathrm{Tr}[(QX^* R^{-1} XQ^{-1})^{-1}].$$

(c) We point out one standard situation in which the lumv estimators are bad, in the sense that although the finite-dimensional lumv estimators exist, their variances approach infinity as the finite-dimensional problem approximates more and more closely the infinite-dimensional one. Suppose $X^* R^{-1} X$ is compact, as it is often in the identification problems we consider (see (6)). Let its (non-negative) eigenvalues be $\mu_1, \mu_2, \ldots$, arranged in order of decreasing magnitude. Now the (finitely many, say N) eigenvalues of $QX^* R^{-1} XQ$, call them $\lambda_1, \lambda_2, \ldots, \lambda_N$ also arranged in decreasing order, are the same as those of $B = (X^* R^{-1} X)^{\frac{1}{2}} Q (X^* R^{-1} X)^{\frac{1}{2}}$. Since $X R^{-1} X$ dominates B, $\mu_1 \geqq \lambda_1, \mu_2 \geqq \lambda_2, \ldots, \mu_N \geqq \lambda_N$. Hence

$$\mathrm{Tr}[(QX^* R^{-1} XQ)^{-1}] = \sum_{i=1}^{N} \frac{1}{\lambda_i} \geqq \sum_{i=1}^{N} \frac{1}{\mu_i}$$

As $N = \dim \mathcal{R}(Q) \to \infty$, $\sum^{N} \frac{1}{\mu_i}$ diverges to $+\infty$, hence $\mathrm{Tr}[(QX^* R^{-1} XQ)^{-1}]$ diverges to $+\infty$.

(d) The chief difficulty in investigating the infinite-dimensional estimators is, of course, that in general R and XX^* do not commute. Further, if one wants to compare finite-dimensional estimators, as for the case of Eq. (8), with the infinite-dimensional ones, P will in general not commute with R or with XQX^*. However, one can set up the artificial problems where these operators do commute, to get some idea of the kinds of behavior the estimators exhibit.

THE MODIFIED ESTIMATORS

Discussion

Since in the system identification application of interest here, as discussed in the introduction, we assume the unknown h lies in a known compact set, there is additional prior information available for the estimation of h beyond what one has usually for an estimation problem as set by Eq. (4). Furthermore, as some previous remarks indicate, the conventional lumv estimation approach is often unsatisfactory - even though the problem is ultimately boiled down to a finite-dimensional one that approximates the original problem. It is fairly obvious that the additional information available can be used to modify lumv estimation, or other standard estimation procedures, so as to yield better estimates, and that this can be done in various ways. A modified lumv estimator is defined below which appears to be satisfactory and for which the mathematical justification is very simple. It is linear, and it can be easily implemented in many applications, at least. It can hardly be called optimal in any meaningful sense, although it does minimize a reasonable bound on mean-squared error.

The Estimator

We suppose that the conditions of (C)2. above are satisfied, so that $\hat{h}_1 = (X^* R^{-1} X)^{-1} X^* R^{-1} y$ is well-defined and is the lumv estimator for h_1. Obviously, then, the original problem can be put in the equivalent form

$$f = h_1 + (X^* R^{-1} X)^{-1} X^* R^{-1} n = h_1 + \zeta \qquad (9)$$

where $f = (X^* R^{-1} X)^{-1} X^* R^{-1} y$, and ζ is defined in context. In this form the lumv estimator for h_1 is simply f. The mean of the new "noise" ζ is zero, and its covariance operator is $(X^* R^{-1} X)^{-1}$. Suppose further that the bounded Hermitian operator $(X^* R^{-1} X)$ has discrete spectrum so that $(X^* R^{-1} X)^{-1}$ does also, and write $(X^* R^{-1} X)^{-1} \psi_i = \sigma_i^2 \psi_i$, $i = 1, 2, \ldots$ with the ψ_i chosen to be orthonormal. The ψ_i span $\mathcal{N}(X)^\perp$, so we can write $\hat{h}_1 = \sum_i \langle f, \psi_i \rangle \psi_i$.

Let a set in $\mathcal{N}(X)^\perp \subset \mathcal{H}$ of the form $\{h = \sum_i a_i \psi_i : |a_i| \leqq b_i > 0, i = 1, 2, \ldots; \sum_i b_i^2 = b < \infty\}$ be called a compact parallelopiped with edges b_i and axes ψ_i. Such a set is compact, and it is trivial to verify that any compact set in $\mathcal{N}(X)^\perp$ is a subset of some compact parallelopiped with axes ψ_i. We now suppose that the set $\mathcal{C} \cap \mathcal{N}(X)^\perp$, which contains the unknown h_1, and which is compact, is contained in a compact parallelopiped $\mathcal{B}$, with edges

$b_1, b_2, \ldots,$ taken in order of decreasing magnitude.

Let us consider an arbitrary linear estimate of h_1: $\hat{h}_1 = \sum_{i,j} a_{ij}(f, \psi_j)\psi_i$ where the a_{ij} are real numbers. Put $\zeta_i = \langle \zeta, \psi_i \rangle$; then $E\zeta_i = 0$, $E\zeta_i\zeta_j = \sigma_i^2 \delta_{ij}$. It is now a routine calculation to show that the mean-squared error is:

$$E\|\hat{h} - h_1\|^2 = \sum_i [(a_{ij} - 1)\langle h_1, \psi_i \rangle + \sum_{j \neq i} a_{ij} \langle h_1, \psi_j \rangle]^2 + \sum_{i,j} a_{ij}^2 \sigma_j^2 \qquad (10)$$

Since the maximum value of $|\langle h_1, \psi_i \rangle| = b_i$, it follows that for all $h_1 \in \mathcal{C} \cap \mathcal{N}(X)^{\perp}$,

$$E\|\hat{h} - h_1\|^2 \leqq \sum_i (|a_{ii} - 1| b_i + \sum_{j \neq i} |a_{ij}| b_j)^2 + \sum_{i,j} a_{ij}^2 \sigma_j^2 \qquad (11)$$

Call this upper bound $M(a)$. The estimator we are seeking, call it $\hat{h}_1^b$, is obtained by choosing the a_{ij} to minimize $M(a)$. This is accomplished by setting $a_{ij} = 0$, $i \neq j$, and $a_{ii} = (b_i^2)/(b_i^2 + \sigma_i^2)$. Thus

$$\hat{h}_1^b = \sum_j \frac{b_j^2}{b_j^2 + \sigma_j^2} \langle f, \psi_j \rangle \psi_j \qquad (12)$$

The mean-squared error is dominated by $M(b, \sigma)$,

$$E\|\hat{h}_1^b - h_1\|^2 = \sum_i \left(\frac{b_i^2}{b_i^2 + \sigma_i^2} - 1 \right)^2 \langle h_1, \psi_i \rangle^2 + \sum_i \left(\frac{b_i^2}{b_i^2 + \sigma_i^2} \right)^2 \sigma_i^2$$

$$\leqq \sum_i \left(\frac{\sigma_i^2}{b_i^2 + \sigma_i^2} \right)^2 b_i^2 + \sum_i \left(\frac{b_i^2}{b_i^2 + \sigma_i^2} \right)^2 \sigma_i^2 = \sum_i \frac{\sigma_i^2 b_i^2}{\sigma_i^2 + b_i^2} \overset{d}{=} M(b, \sigma). \qquad (13)$$

Comments:

(a) $\hat{h}_1^b$ is, of course, a biased estimator - a fact of very little significance in the applications envisioned in this paper.

(b) The corresponding lumv estimator $\hat{h}_1$ has variance $= E\|\hat{h}_1 - h_1\|^2 = \sum_i \sigma_i^2$. $M(b, \sigma)$, the bound on $E\|\hat{h}_1^b - h_1\|^2$, is always less than this value. As the b_i become large with respect to the σ_i, $\hat{h}_1^b$ approaches the lumv estimator.

(c) The estimates given by $\hat{h}_1^b$ can perfectly well violate the condition that h_1 lies in the compact set $\mathcal{C} \cap \mathcal{N}(X)^{\perp}$. Usually this would not be very important, since in practice the specification of $\mathcal{C}$ is likely to be fuzzy. If it is important, one can of course define the estimator

$$\tilde{h}_1^b \overset{d}{=} \sum_n \min\left[\frac{b_n^2}{b_n^2 + \sigma_n^2} |\langle f, \psi_n \rangle|, b_n \right] \operatorname{sgn} \langle f, \psi_n \rangle \psi_n$$

$\tilde{h}_1^b$ has uniformly smaller mean-squared error than $\hat{h}_1^b$, but it is nonlinear, and more difficult to implement.

Example

Let a system be represented by the equation

$$y(t) = \int_A h_1(t,s)x(s)ds + \iint_{AA} h_2(t,s_1,s_2)x(s_1)x(s_2)ds_1ds_2 + n(t),\ t \in T \qquad (14)$$

where A and T are finite intervals. The input x must belong to $L_2(A)$. The kernels h_1 and h_2 are to be square-integrable over $T \times A$ and $T \times A \times A$, respectively. The transformation H is characterized by the pair of kernels $h = (h_1, h_2)$. We regard h as an element of $\mathcal{H} = L_2(T \times A) \oplus L_2(T \times A^2)$, so $\|h\|^2 = \|h_1\|^2 + \|h_2\|^2$, where the norms on the right side are Hilbert-space norms. The unknown h is specified as belonging to $\mathcal{C}$, a compact subset of $\mathcal{H}$.

Let $\{\eta_n(t)\}$ be a complete orthonormal system (cons) in $L_2(T)$ and $\{\phi_n(t)\}$ a cons in $L_2(A)$. We write: $y(t) = \sum_i c_i \eta_i(t)$, $x(s) = \sum_i a_i \phi_i(s)$ and $h_1(t,s) = \sum_{i,j} b_{ij}\eta_i(t)\phi_j(s)$, $h_2(t,s_1,s_2) = \sum_{i,j,k} b_{ijk}\eta_i(t)\phi_j(s_1)\phi_k(s_2)$ where of course there is an abuse of notation, because the series all converge in the appropriate L_2 norm rather than pointwise. The fact that $\mathcal{C}$ is compact implies that there exist positive integers M and N such that if h_1 is approximated by $\sum_{i=1}^{M}\sum_{j=1}^{N} b_{ij}\eta_i(t)\phi_j(s)$ and h_2 similarly by $\sum_{i=1}^{M}\sum_{j,k=1}^{N} b_{ijk}\eta_i(t)\phi_j(s_1)\phi_k(s_2)$ then h is approximated uniformly to within ϵ in norm. Using these expansions, Eq. (13) is approximated by M linear algebraic equations in the unknowns b_{ik} and b_{ikj} (see (6)). Let P_ϵ be the orthogonal projection onto the span of $\{\eta_1, \ldots, \eta_M\}$, and let us now suppose that $P_\epsilon n$ is a white noise - more precisely suppose that the random variables $<n, \eta_i>$, $i = 1, \ldots, M$, are uncorrelated with mean zero and variance σ^2.

Let the problem be to determine h as well as possible, with a one-shot procedure, i.e., a single input signal $x = \sum^{N} a_n \phi_n$ is used and one test is made. For convenience, suppose x is chosen to have norm 1. It is readily verified that the operator $X^* R^{-1} X = \frac{1}{\sigma^2} X^* X$ is defined by the equation

$$X^*Xh = \begin{bmatrix} x(s)\int x(s')[\cdot]ds' & , \quad x(s)\iint x(s_1')x(s_2')[\cdot]ds_1'ds_2' \\ x(s_1)x(s_2)\int x(s')[\cdot]ds', & x(s_1)x(s_2)\iint x(s_1')x(s_2')[\cdot]ds_1'ds_2' \end{bmatrix} \begin{bmatrix} h_1(t,s') \\ h_2(t,s_1',s_2') \end{bmatrix} \qquad (15)$$

where the formal rules of matrix multiplication are followed, except h_1 and h_2 are inserted in the integrals as indicated by $[\cdot]$. In terms of the η_i and

ϕ_j this is readily modified to give an expression for X^*PX. Finally, it can again easily be verified that the orthonormal eigenvectors of $X^*PR^{-1}PX = \frac{1}{\sigma^2}X^*PX$ are the pairs, written as column vectors,

$$\psi_p(t,s,s_1,s_2) = \frac{1}{\sqrt{2}}\begin{bmatrix} \eta_p(t)\sum_{\mu=1}^{N} a_\mu \phi_\mu(s) \\ \\ \eta_p(t)\sum_{\nu=1}^{N} a_\nu \phi_\nu(s_1) \quad \sum_{\rho=1}^{N} a_\rho \phi_\rho(s_2) \end{bmatrix}, \quad p = 1,\ldots,M, \qquad (16)$$

and the eigenvalues are all equal to $\frac{2}{\sigma^2}$. Thus the σ_p^2, the eigenvalues of $(X^*PR^{-1}PX)^{-1}$ are all $\sigma^2/2$.

To find the estimator given by Eq. (12) we observe first, using the fact that $X^*\eta_n = \sqrt{2}\,\psi_n$, $f = (X^*PR^{-1}X^*PR^{-1}\sum_{n=1}^{\infty} c_n\eta_n = (X^*PX)^{-1}X^*\sum_{n=1}^{M} c_n\eta_n = \frac{\sqrt{2}}{2}\sum_{n=1}^{M} c_n\psi_n$. Then,

$$\hat{h}_1^{b} = \sum_{n=1}^{M} \frac{b_n^2}{b_n^2 + \frac{\sigma^2}{2}} \, \frac{\sqrt{2}}{2} \, c_n\psi_n = \sum_{n=1}^{M} \frac{\sqrt{2}\, b_n^2}{2b_n^2 + \sigma^2} \, c_n\psi_n \qquad (17)$$

Of course h is still largely undetermined, because in a problem even of this complexity $\mathcal{N}(P_\epsilon X)$ is large when X is the mapping correspond to one-shot identification. It is straightforward to carry out this example for m measurements (m-shot identification), made with m possibly different x's, but it will not be done here. We remark simply that the operator X^*X for the m-shot identification is $X_1^*X_1 + \cdots + X_m^*X_m$ where the X_i correspond to the i'th imput signal x_i. There are cases in which determining the eigenvalues and eigenvectors of X^*X is simple, as, e.g., when all the x_i are equal, but in general it requires a non-trivial calculation.

REFERENCES

(1). "On the Measurement and Use of Time-Varying Communication Channels," W. L. Root, Info. and Control 8(1965) pp 390-422.

(2). "Determinable Classes of Channels," R. T. Prosser and W. L. Root, Proc. Symp. on Syst. Theory, Poly. Inst. of Brooklyn (1965), pp 133-157.

(3). "On the Identification of Unknown Systems," R. T. Prosser and W. L. Root, Proc. First Princeton Conf. on Info. Sciences and Systems (1967) pp 285-289.

(4). "On Determinable Classes of Signals and Linear Channels," P. P. Varaiya, SIAM J. Appl. Math., 15 (1967) pp 440-449.

(5). "Bounded Determinable Subsets of Banach Spaces," B. S. Rajput, SIAM J. Appl. Math., 20 (1971) pp 735-748.

(6). "On The Structure of a Class of System Identification Problems," W. L. Root, Automatica, 7(1971) pp 219-231.

(7). "Some General Structure Theory of Systems to be Used in Identification and Measurement," W.L. Root, 5th Princeton Conf. on Info. Sci. &Sys. (1971).

(8). "Introduction to Optimization Theory in Hilbert Space," A V. Balakrishnan, Lecture Notes, Springer-Verlag, Berlin, (1971).

OPTIMAL AND SUBOPTIMAL RECURSIVE NONLINEAR FILTERING

H. W. SORENSON

University of California at San Diego, La Jolla, California 92037

ABSTRACT

The general problem of recursively estimating the state of a nonlinear, dynamical system from noisy measurement data is reviewed and approaches currently being investigated for its solution are discussed and compared. In particular, the extended Kalman filter and its variants are compared with techniques which attempt to compute the a posteriori density function in order to obtain "optimal" estimates.

1. THE RECURSIVE NONLINEAR FILTERING PROBLEM

Attention here is directed exclusively to systems which are observed at discrete time instants. Although more general models can be formulated, it is convenient for the following discussion to consider systems with plant and measurements described by

$$\underline{x}_{k+1} = \underline{f}_k(\underline{x}_k) + \underline{w}_k \quad ; \quad k = 0, 1, 2, \ldots \tag{1.1}$$

$$\underline{z}_k = \underline{h}_k(\underline{x}_k) + \underline{v}_k \quad ; \quad k = 0, 1, 2, \ldots \tag{1.2}$$

where the subscript k denotes the k^{th} measurement time. The state $\underline{x}$ is n-dimensional and the measurement vector $\underline{z}$ is m-dimensional. The plant and measurement noise sequences* $\underline{w}^k$ and $\underline{v}^k$ are assumed to be zero mean, white gaussian sequences with autocorrelation matrices

$$E[\underline{w}_k \underline{w}_j^T] = Q_k \delta_{kj} \quad ; \quad E[\underline{v}_k \underline{v}_j^T] = R_k \delta_{kj}$$

The noise sequences are mutually independent and are independent of the state. The initial state is assumed to have a known probability density function $p(\underline{x}_0)$.

In the absence of plant and measurement noise, the problem that is considered below would have the following simple, deterministic statement. Determine the state $\underline{x}_k$ from the measurement data $\underline{z}^k$. As will be discussed subsequently, the solution of this problem is intimately connected with the question of system observability. When stochastic effects are included in the model as in (1.1)-(1.2), the statement of the filtering problem comes to include an element of arbitrariness. Certainly the basic objective is to "estimate" the state from the data. However, the random effects in the system model implies that redundant data must be collected in order to minimize the noise influence on the estimate. Now, it becomes reasonable to define "best" estimates which introduces the arbitrariness mentioned immediately above.

A widely-accepted definition of the filtering problem states that the data are to be used to determine the a posteriori density function $p(\underline{x}_k | \underline{z}^k)$. This density function provides the most complete description of the system and the information provided by the data about the system that is possible. All subsequent discussion will be addressed to the problem of determining or approximating $p(\underline{x}_k | \underline{z}^k)$ and will be referred to as the filtering problem.

---

* The superscript notation is used to denote a collection of vectors. That is, $\underline{a}^k = (\underline{a}_0, \underline{a}_1, \ldots, \underline{a}_k)$.

While the a posteriori density provides a complete solution of the filtering problem, it has the disadvantage that it is a function rather than a finite-dimensional estimate. If the problem were deterministic, the solution would be provided by the vector $\hat{\underline{x}}_{k|k}$ that satisfied the plant and measurement equations for all k. A similar "solution" is commonly sought for the stochastic problem. To obtain estimates $\hat{\underline{x}}_{k|k}$ criteria are defined which lead to optimal estimates. Two of the more important criteria are the following.

Minimum Mean-square Estimates: The estimate $\hat{\underline{x}}^{MS}_{k|k}$ of the state $\underline{x}_k$ based on the measurement data $\underline{z}^k$ is chosen so that the mean-square error $E[(\underline{x}_k - \hat{\underline{x}}_{k|k})^T (\underline{x}_k - \hat{\underline{x}}_{k|k})]$ is minimized. The estimate that accomplishes the minimization is provided by the conditional mean

$$\underline{x}^{MS}_{k|k} = E[\underline{x}_k | \underline{z}^k] \tag{1.3}$$

Maximum A Posteriori Estimates: The estimate $\hat{\underline{x}}^{MAP}_{k|k}$ of the state $\underline{x}_k$ based on the data $\underline{z}^k$ is chosen so that the a posteriori density is maximized.

$$p(\hat{\underline{x}}^{MAP}_{k|k} | \underline{z}^k) = \max_{\underline{x}_k} p(\underline{x}_k | \underline{z}^k) \tag{1.4}$$

Note that both estimates can be determined if the a posteriori density is known.

Much of the current research on nonlinear filtering is concerned with recursive formulations in which the solution for the $(k-1)^{th}$ stage is used to obtain the solution for the k^{th} stage. Only the recursive formulation shall be considered here. A general solution of the recursive filtering problem can be obtained through Bayes' rule. It is not difficult to show [1] that the a posteriori density evolves in the following manner.

Bayesian Recursion Relations

$$p(\underline{x}_k | \underline{z}^k) = c_k \, p(\underline{x}_k | \underline{z}^{k-1}) \, p(\underline{z}_k | \underline{x}_k) \tag{1.5}$$

$$p(\underline{x}_k | \underline{z}^{k-1}) = \int p(\underline{x}_{k-1} | \underline{z}^{k-1}) \, p(\underline{x}_k | \underline{x}_{k-1}) \, d\underline{x}_{k-1} \tag{1.6}$$

where the normalizing constant c_k is

$$1/c_k = p(\underline{z}_k | \underline{z}^{k-1}) = \int p(\underline{x}_k | \underline{z}^{k-1}) \, p(\underline{z}_k | \underline{x}_k) \, d\underline{x}_k$$

and the initial condition for (1.5) is

$$p(\underline{x}_0 | \underline{z}^{-1}) = p(\underline{x}_0)$$

The densities $p(\underline{z}_k | \underline{x}_k)$ and $p(\underline{x}_k | \underline{x}_{k-1})$ can be written more explicitly by introducing the a priori assumptions about the system. It is seen, assuming Q_k^{-1} and R_k^{-1} exist, that

$$p(\underline{z}_k | \underline{x}_k) = k_v \exp\{-\tfrac{1}{2}[\underline{z}_k - \underline{h}_k(\underline{x}_k)]^T R_k^{-1} [\underline{z}_k - \underline{h}_k(\underline{x}_k)]\} \tag{1.7}$$

$$p(\underline{x}_{k+1} | \underline{x}_k) = k_w \exp\{-\tfrac{1}{2}[\underline{x}_{k+1} - \underline{f}_k(\underline{x}_k)]^T Q_k^{-1} [\underline{x}_{k+1} - \underline{f}_k(\underline{x}_k)]\} \tag{1.8}$$

Utilization of (1.7) and (1.8) in conjunction with the prescribed initial condition $p(\underline{x}_0)$ provides the information required for the determination of $p(\underline{x}_k | \underline{z}^k)$ for any k through evaluation of (1.5)-(1.6). Thus, a general solution of the nonlinear filtering problem is available. Unfortunately, the actual evaluation of the Bayesian recursion relations for a specific nonlinear system is not accomplished in a trivial manner. It is to the problem of developing computational algorithms for evaluating (1.5)-(1.6) for specific systems that the remainder of this discussion is directed.

2. APPROXIMATE SOLUTION OF THE BAYESIAN RECURSION RELATIONS

Virtually the only general system for which closed-form solutions of (1.5)-(1.6) can be found is when the plant and measurement equations are linear. Then, the a posteriori density is gaussian and the conditional mean and covariance are described by the Kalman filter equations [2, 3, 4].

Considerations important to the development of computational algorithms for the evaluation of (1.5) and (1.6) are the following.

(1) The a posteriori density is a function of the measurement data $\underline{z}^k$. A computational algorithm must provide for the efficient representation of this function in order to permit a practicable implementation. Note that in the linear, gaussian case this is accomplished through the relations defining the conditional mean and covariance which for this special case are sufficient statistics. This provides a very succinct representation of the density.

(2) The calculation of the filtering density $p(\underline{x}_k | \underline{z}^k)$ using (1.5) is seen to require the multiplication of two density functions. This does not represent a difficult task other than in the storage requirements that are implied in such an operation. However, the normalization and the determination of moments requires integration of the product and can introduce a considerable burden for an algorithm.

(3) The determination of the prediction density $p(\underline{x}_{k+1} | \underline{z}^k)$ can be viewed as involving two major operations. First, a nonlinear transformation of the random variable $\underline{x}_k$ to $\underline{f}_k(\underline{x}_k)$ must be accomplished. If no plant noise exists, the density of the transformed variable is equal to the prediction density. When plant noise exists, the addition of the independent random variables $\underline{f}_k$ and $\underline{w}_k$ implies that a convolution must be performed in order to obtain $p(\underline{x}_{k+1} | \underline{z}^k)$. These operations impose a substantial burden on any procedure. Note that the transformation and convolution are performed simultaneously through direct evaluation of (1.6).

With these three points in mind, consider some of the procedures that have been devised for recursive nonlinear filtering. These procedures can be regarded as forming two basic classes. First, there are procedures that are based on the Kalman filter equations. These methods involve expansions relative to a prescribed reference state and shall be referred to as local nonlinear filtering methods. Second, the past two or three years have seen the development of procedures which attempt to approximate $p(\underline{x}_k | \underline{z}^k)$ everywhere in state space. This class shall be referred to as global nonlinear filtering methods.

2.1 Local Nonlinear Filtering Methods

Virtually the only recursive nonlinear filtering method that has seen application to practical problems is the so-called extended Kalman filter and its variants which include iterative and/or divergence control logic.

Extended Kalman Filter Equation [5]

For the system (1.1)-(1.2), the basic extended Kalman filter equations have the following form.

$$\hat{\underline{x}}_{k|k-1} = \underline{f}_k(\hat{\underline{x}}_{k-1|k-1}) \tag{2.1a}$$

$$P_{k|k-1} = \Phi_{k,k-1} P_{k-1|k-1} \Phi_{k,k-1}^T + Q_{k-1} \tag{2.1b}$$

$$\hat{\underline{x}}_{k|k} = \hat{\underline{x}}_{k|k-1} + K_k[\underline{z}_k - \underline{h}_k(\hat{\underline{x}}_{k|k-1})] \tag{2.1c}$$

$$P_{k|k} = P_{k|k-1} - K_k H_k P_{k|k-1} \tag{2.1d}$$

$$K_k = P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + Q_k)^{-1} \tag{2.1e}$$

where $\Phi_{k,k-1}$ and H_k represent the partial derivative matrices

$$\Phi_{k,k-1} = \frac{\partial \underline{f}_{k-1}}{\partial \underline{x}_{k-1}}(\hat{\underline{x}}_{k-1|k-1}) \; ; \quad H_k = \frac{\partial \underline{h}_k}{\partial \underline{x}_k}(\hat{\underline{x}}_{k|k-1})$$

evaluated with $\hat{\underline{x}}_{k-1|k-1}$ and $\hat{\underline{x}}_{k|k-1}$ respectively. Note that the nonlinear equations are used to obtain the predicted estimate $\hat{\underline{x}}_{k|k-1}$ and to form the residual $[\underline{z}_k - \underline{h}_k(\hat{\underline{x}}_{k|k-1})]$.

The adequacy of this procedure depends strongly on the validity of the linear approximations used to obtain $\Phi_{k,k-1}$ and H_k. As is well-known, practical experience has shown that such approximations are often valid and this procedure has produced satisfactory results. A problem called divergence [3, 5, 6] has been encountered, not infrequently however, that appears to mar the general success of the method. Essentially, the error in the estimates become larger than justified by the error covariance matrix $P_{k|k}$ and indicate that the filter is producing unreliable and possibly misleading results. When this results from errors in the linear approximations, it is often possible to overcome divergence by iterating at each stage. For example, data can be processed at stage k as indicated in (2.1) but taking care to save $P_{k|k-1}$. Then, the estimate $\hat{\underline{x}}_{k|k}$ can be used to relinearize the measurement equation. Using $\hat{\underline{x}}_{k|k}$ as $\hat{\underline{x}}_{k|k-1}$ (2.1c)-(2.1e) can be reevaluated. The procedure can be repeated until no change is observed. The procedure could be extended by using the estimate $\hat{\underline{x}}_{k|k}$ to compute and smoothed estimate $\hat{\underline{x}}_{k-1|k}$. Then, the entire cycle could be repeated including the prediction [7].

Divergence often occurs because the model does not adequately describe the system. Then, methods have been devised which cause the error covariance matrix to be increased. Generally this is accomplished by increasing Q_{k-1} or R_k, [3, 5] by multiplying $P_{k|k-1}$ by a constant [6], or by restarting the filter periodically [5]. The practical implementation [6] of these procedures generally is based on the observation of the residual sequence and the computation of sample averages. If the linear assumptions are reasonably valid, the residual sequence should be white noise with covariance computed from the filter. The filter behavior is generally judged to be satisfactory if the low frequency content of the residual is eliminated and the sample variance is consistent with its theoretical value.

The use of the extended Kalman filter implies that the a posteriori density $p(\underline{x}_k|\underline{z}^k)$ is gaussian. As will be seen in Section 3, this could be an extremely poor approximation of the density function if all possible values of the state must be considered. However, depending upon the quality of the a priori information, the gaussian approximation may be quite satisfactory in a region around the estimate. This statement will be reconsidered in Section 3.

The use of the extended Kalman filter involves a local approximation whose range of validity must obviously be limited. Techniques which attempt to extend the range of validity of the local approximation have been proposed. Most prominent among these have been methods [7, 8] which include second-order terms in the Taylor series expapnsion of $\underline{f}_k$ and $\underline{h}_k$. In these methods, the assumption that $p(\underline{x}_k|\underline{z}^k)$ is gaussian is retained explicitly in order to overcome the "moment" problem.

Since the objective of second-order methods is to extend the range of the local approximation but fail to give up the basic gaussian assumption, it is not surprising that numerical studies indicate that iterative techniques are generally as effective [7]. Since the Kalman filter equations are easily adapted to iterate and since second-order filters have substantially more complicated form and require more computation per cycle, the practicality of the latter are suspect at best.

One of the first serious attempts to eliminate the gaussian assumption involved the use of Gram-Charlier or Edgeworth expansions of the density function [9, 10]. This expansion is a series of polynomials orthogonal with respect to the gaussian distribution which can be used as a representation of a large class of density functions. As noted above, the Bayesian relations can be solved in closed-form when gaussian densities are involved. The presence of the gaussian weighting function in the Gram-Charlier suggests that densities with this representation may be more amenable to computation and certainly stimulated interest in this expansion. Unfortunately, this initial utilization of the Gram-Charlier expansion suffered from the failing that it was based on a local theory. That is, the coefficients of the series were determined using expansions about a prescribed reference point. As a consequence, a large number of terms were required to obtain a reasonable approximation of a distinctly nongaussian a posteriori density so its utility is severely limited. On the other hand, these studies indicated that for cases in which only unimodal distributions had to be considered the approach did provide improved local approximations. However, the mean and variance of the resulting density were seldom substantially different than those obtained by second order or iterative methods. This behavior suggests that it will be a rare case in which a nongaussian density approximation will be required when the a priori information and data permits the consideration of a single mode of the a posteriori density.

2.2 Global Nonlinear Filtering Methods

During the past two or three years, several methods which approximate the a posteriori density over the entire region containing significant probability mass have been proposed. These eliminate the restriction to gaussian approximations and avoid the local approximations discussed above. These approaches will be outlined below.

FLOATING GRID: Bucy [11] in conjunction with Senne [12] has defined and investigated an approximation procedure which they have referred to as a floating grid. A grid of points is defined in state space to which finite probability mass is assigned. The amount of mass is determined from the a priori density. The representation of the density by a finite number of points provides a straightforward means of performing the convolution and the integrations required to determine moments. A substantantial storage burden is imposed by this approximation. To reduce the storage requirements several techniques have been developed which have resulted in substantial computational savings relative to "brute-force" application of this point-mass approximation. Nonetheless, Bucy and Senne [12] state that the dimension of the state space must be no greater than four for the computation time per estimate to be less than one second on a machine with parallel processing capability and basic machine cycle time equal to that of the CDC 6600. This observation clearly indicates the fundamental difficulty which arises in attempting to accomplish global nonlinear filtering. That is, the computational burden implicit in global filtering is generally enormous relative to extended Kalman filtering and related local filtering procedures.

GAUSSIAN SUMS: Alspach and Sorenson [13, 14] have proposed to approximate the a posteriori density function by a weighted sum of gaussian density functions. This approach is similar to the floating grid approach in the sense that a grid is introduced which defines the mean-value of each gaussian term in the series. As in the floating grid approach the approximation converges uniformly to the true density as the number of grid points increases without bound.

For the gaussian sum filter the a posteriori density at each stage is approximated by a weighted sum of gaussian density functions. For example, a density p is represented as[*]

$$p(\underline{x}) = \sum_{i=1}^{q} \alpha_i N(\underline{x}-\hat{\underline{x}}_i, P_i)$$

where $\alpha_i \geq 0$ for all i and $\sum_{i=1}^{q} \alpha_i = 1$. The moments $\hat{\underline{x}}_i, P_i$ are up-dated using extended Kalman filter equations so that the resulting filter has the characteristic that the conditional mean and covariance are obtained as a weighted sum of extended Kalman filters. The details of the approximation procedure are discussed in detail in Reference 13. The numerical results presented in Section 3 are based on the gaussian sum approximation.

Recent papers by Center [15] and by Tse [16] point out the similarities of the floating grid and gaussian sum approaches and suggest other possible approximations. Tse [16], in particular, discusses the use of parallel processing computers that may make these approaches computationally practical in the future.

SPLINE FILTERS: de Figuiredo [17] has proposed the use of multivariate fundamental splines for calculation of the a posteriori density. This recent work provides an interesting alternative approach to the global nonlinear filtering problem.

3. NUMERICAL RESULTS FOR NONLINEAR EXAMPLE

Consider the following scalar example

$$x_{k+1} = x_k + \beta x_k^2 + w_k \tag{3.1}$$

$$z_k = x_k^2 + v_k \tag{3.2}$$

---

[*] $N(\underline{x} - \underline{a}, B) \triangleq k \exp \left\{ -\frac{1}{2} (\underline{x} - \underline{a})^T B^{-1}(\underline{x} - \underline{a}) \right\}$

where it is convenient to assume that the noise sequences are stationary so that the variances are the same for all k. Let

$$\sigma_w^2 = Q_k \quad ; \quad \sigma_v^2 = R_k .$$

Also, suppose that the initial state is gaussian with

$$E[x_0] = 1 \quad ; \quad E[(x_0 - 1)^2] = 1$$

This example, while simple, exhibits many of the basic considerations required for more complicated systems.

An important consideration for linear filtering applications relates to the system observability. Similar considerations are important for nonlinear systems but a concise global solution of the general problem, such as exists for linear systems [18], has not been discovered. Recent papers by Fitts [19] and Griffith and Kumar [20] provides some results for the nonlinear observability problem. For discrete-time systems the system

$$\underline{x}_{k+1} = \underline{f}_k(\underline{x}_k)$$

$$\underline{z}_k = \underline{h}_k(\underline{x}_k)$$

is said to be q-stage observable if the initial state $\underline{x}_0$ can be uniquely determined from q measurements $\underline{z}^{q-1}$. The requirement that the initial state be determined uniquely implies that for any two different initial states, the two output sequences cannot be identical. System observability is a fundamental requirement for any filtering application. If the system is observable, then for sufficiently large signal-to-noise ratios, the a posteriori density will become unimodal. As demonstrated below, q-stage observability does not insure that the a posteriori density will become unimodal in q-stages. Stochastic influences can delay the convergence to a unimodal condition.

The scalar system (3.1)-(3.2) is easily seen to be unobservable for any q if $\beta = 0$. In this case, it is impossible to distinguish the sign of the state. However, if $\beta \neq 0$, the system is 2-stage observable. These properties manifest themselves in the numerical results presented below.

Typical results for several filters are presented and discussed below. In particular, the gaussian sum filter, the extended Kalman filter, the extended Kalman filter with iteration, and a second-order filter [8] are applied to several situations. In all of these studies it was found that the gaussian sum approximation produced results that are indistinguishable from the true density. Thus, the gaussian sum results are treated as the exact a posteriori density.

The actual value of the initial state, the plant noise variance σ_w^2, the measurement noise variance σ_v^2, and the plant nonlinearity parameter β are the only parameters of the problem. No Monte Carlo results are presented here but can be found in Reference 21. Instead, typical results for single trajectories and measurement sequences are shown.

Consider, first, cases in which no plant nonlinearity exists (i. e. $\beta = 0$). Since the system is unobservable, suppose that x_0 is initially positive and that the filter is given this information. Then, it is possible to consider only the positive solution of the measurement equation.

The behavior of the four different filters is strongly dependent upon the measurement signal-to-noise ratio. For example, when $x_0 = 0.5$, $\sigma_w^2 = 0$, $\sigma_v^2 = 0.05$ and $\beta = 0$, the behavior of the three local filters is statistically undistinguishable. That is, the estimation errors observed for the extended Kalman filter, the iterated filter, and the second-order filter are essentially the same. Also, in each case, the approximating gaussian a posteriori density follows the positive peak of the true a posteriori density. However, when the initial state is reduced to $x_0 = +0.2$, differences are observed. A typical result for 16 measurement samples is depicted in Figure 1. The second-order filter estimates the state to be equal to zero but has a very large error variance. This behavior can be shown analytically to be a property of the second-order filter for this system. Thus, it essentially provides an estimate that is the average of the two possible solutions and indicates its uncertainty by a large variance. Thus, its behavior is unsatisfactory. The extended Kalman filter can be seen to have diverged in the sense that it provides an estimate containing a substantial bias. However, the iterated filter eliminates the bias and is observed to correspond with the positive peak of the a posteriori density. It should be noted in this case that the residual for the iterated filter exhibits very little low-frequency content and has a variation that is consistent with the gaussian a posteriori approximation implied by the filter equations.

When plant noise with variance $\sigma_w^2 = 0.1$ is introduced into the system very different results are obtained. Certainly, the state changes from one stage to the next. However, the extended Kalman filter produces results that are substantially the same as the results obtained when an iteration is utilized. Thus, the presence of plant noise has prevented the gain from becoming too small and thereby improved filter performance. The second-order filter again performed poorly although its error variance was large and correctly indicated poor filter behavior. The true a posteriori and the gaussian approximation indicated by the iterated filter are shown for sixteen measurements in Figure 2. The satisfactory performance of the iterated filter is manifest.

Finally, consider the behavior of the density when a plant nonlinearity is present so that the system is observable. Let $x_0 = -0.2$, $\beta = -0.2$, $\sigma_w^2 = 0$, and $\sigma_v^2 = 0.05$. The behavior of the local filters depends entirely on the choice of the initial estimate. If the initial state state estimate for the local filters is assigned a negative value approximately equal to -0.2, the linear filters can be made to perform satisfactorily. However, a positive initial estimate leads to unsatisfactory results. Consequently, only the true a posteriori density is shown in Figure 3.

The behavior of the a posteriori density is interesting. Recall that the initial density has mean value and variance both equal to one. Thus, this information causes the positive peak to be largest for several samples. Finally, at sample 10 the negative peak becomes largest but then reverts to the smaller at sample 11. From sample 12 to 15 the positive peak diminishes rapidly, however, and vanishes to all practical purposes following sample 16. It should be noted that this behavior depends upon the measurement noise sequence but is typical of possible density behavior. Thus, even though the system is 2-stage observable, it is not until the 16th sample that a unimodal density is obtained. (It remains unimodal thereafter.) If, for example, x_0 is set equal to -0.5, the unimodality appears much more quickly. This demonstrates the influence that the signal-to-noise ratio can have upon the length of this transient phase.

In Figure 3 the negative peak of the density function can be seen to be located consistently near the true value of the state. Thus, one would expect that the residual formed using the MAP estimates would exhibit random behavior not unlike that which occurs for linear gaussian system. In fact, a suitably normalized gaussian approximation of the single peak centered on the estimate could possibly be used for a consistency test for the residual.

Observe from Figure 3 that the minimum mean-square estimate $\hat{x}_{k|k}^{MS}$ will be located between the two peaks and, consequently, contains significant errors until the density approaches unimodality. A maximum a posteriori estimate $\hat{x}_{k|k}^{MAP}$ could conceivably encounter considerable difficulty. For example, it would probably be difficult for any computational algorithm that tracks only the larger peak to switch from the positive to the negative to the positive, and, finally, to the negative peak as is required in Figure 3. It would be more reasonable to develop MAP estimators that track each peak separately until the steady-state behavior is obtained.

4. NONLINEAR FILTERING - A CRITICAL LOOK

Global nonlinear filtering is growing beyond its infancy. As must be true for any infant, the first steps, as exemplified by the work mentioned above, are exhilarating for those involved and can easily lead to overly ambitious claims and unwarranted optimism. Viewed with even a modicum of perspective, however, it becomes obvious that much work remains before the infant will grow to maturity. It is fun and hopefully, worthwhile to attempt to predict the character of the mature development and to suggest some activities that are required to shape the development.

The basic objective of global nonlinear filtering might be regarded as the development of a practical computational algorithm which will permit the determination of the a posteriori density to any prescribed accuracy for any system that can be described, for example, by (1.1) and (1.2). This is the achievement of the Kalman filter for linear, gaussian systems; if it can be accomplished for nonlinear nongaussian systems, the achievement would be worthy of any of the scientific titans of history. The developments described above in Section 2.2 do provide procedures for computing the a posteriori density for any system. But they have the practical limitation that the computational requirements associated with their implementation are enormous. Thus, unless some new insights occur which permit the development of substantially more efficient algorithms, a computer revolution, a not unthinkable possibility, will be necessary to allow the application of current algorithms and their future descendents to practical systems.

Assuming that global nonlinear filtering methods will continue to require substantially more computation than local filtering techniques, it is natural to ask and attempt to answer the following question. Under what conditions is it desirable and necessary to assume the additional computational burden and utilize global nonlinear filtering techniques? Certainly, no answer to this question that could be universally accepted exists at this time. However, some related considerations can be discussed.

Local filtering techniques in general and the extended Kalman filter (hereafter referred to as EKF) in particular are looked upon with scorn in some quarters because these approaches are "suboptimal." In addition, the degree of suboptimality is not readily determined. As a consequence, it is reasonable to solve the global filtering problem if only to provide a reference against which local methods can be compared. However, the continued use of the EKF must be tolerated because it has proven to provide satisfactory results for many nonlinear systems. This is especially true when the filter is designed to monitor the residuals and to initiate corrective action whenever a low frequency component is observed that implies the onset of divergence.

The success of the EKF forces a search for general circumstances in which this local filtering method cannot be expected to perform satisfactorily. Certainly, one of the most important requirements is that an a priori estimate be available which permits the local approximation to be valid initially. If it is impossible to define an appropriate a priori estimate, then the EKF is doomed to failure and a global filter is required. For many systems of interest, this would appear to be an unlikely situation. Frequently, the signal-to-noise ratio is sufficiently large that a reasonable estimate can be obtained using only deterministic models. When more than one solution is possible, physical considerations may permit the determination of the only reasonable solution which can then be used to initiate the EKF. If more than one a priori estimate must be considered, the a posteriori density will be multimodal so the EKF cannot be used.

If the a posteriori density can be regarded as unimodal but nongaussian, the EKF must produce suboptimal results. Thus, it may be desired to utilize local or global procedures which eliminate the gaussian assumption. In many cases, the EKF can be expected to provide pessimistic results since the gaussian density maximizes entropy. As long as the residual is forced to be white, the EKF should produce results that are satisfactory in some ways. More complicated procedures may provide improvements but this would seem to be very problem-dependent.

Finally, the signal-to-noise ratio may be so small that linearizations provide inadequate approximations so that the EKF produces little data filtering. That is, the divergence control logic may require past data to be discounted so strongly that only current data is used in determining the estimate. Then, the estimation error will be comparable or greater than the measurement noise indicating the lack of any filtering (noise removal) activity. In this case global nonlinear filters may be required in order to extract the maximum information from the data.

The preceding discussion suggests some advantages that may accrue from the use of global nonlinear filtering.

(1) The calculation of $p(\underline{x}_k | \underline{z}^k)$ does not depend upon a priori information about the state. Thus, multimodal situations are handled in a straightforward manner.

(2) The gaussian assumption is eliminated so that substantial improvements in estimation accuracy are possible. A simple example demonstrating the improvement possible is presented in Reference 21.

(3) System observability does not have to be investigated prior to determining $p(\underline{x}_k | \underline{z}^k)$. In fact, the behavior of the a posteriori density could be used to accomplish a observability analysis. Of course, an unobservable or marginally observable system might present special computational problems for an algorithm and thereby negate the preceding remarks. Observability considerations have a strong influence on the behavior of the Kalman filter

error covariance matrix.

(4) Finally, calculation of $p(\underline{x}_k | \underline{z}^k)$ provides a reference which can be used to measure the performance of all suboptimal procedures. It bears mentioning that having $p(\underline{x}_k | \underline{z}^k)$ does not necessarily provide the desired information in a readily available form. For one or two-dimensional problems, graphical displays provide immediate insights into the behavior of nonlinear filters. For higher dimensional problems such displays are not possible so analytical tools must be employed to determine information such as mode locations, moments, observability, etc. Depending upon the density representation, nontrivial problems may be encountered which increase the computational burden.

To obtain more precise answers to the basic question raised above, considerably more computational experience with specific problems and algorithms is needed. Paramount in the meaningful development of such experience is the definition of test problems that can be used to compare algorithms. Problems that would be useful must necessarily provide special challenges which indicate the strengths and weaknesses of specific algorithms and which assist in the quantitative definition of conditions which require the use of specific types of procedures. Several problems have appeared in the literature (e.g. [7], [9], [12], [14], [22], [23], [24]) which can serve as candidates as standard test problems.

REFERENCES

1. Ho, Y. C., Lee, R. C. K., "A Bayesian Approach to Problems in Stochastic Estimation and Control," Proc. JACC, 1964, pp. 382-387.

2. Kalman, R. E. "A New Approach to Linear Filtering and Prediction Problems," J. Basic Engr., 82D, March 1962, pp. 35-45.

3. Sorenson, H. W., "Kalman Filtering Techniques," Advances in Control Systems, Vol. 3, Chapt. 5, Academic Press, N. Y., 1966.

4. Sorenson, H. W., "Comparison of Kalman, Bayesian, and Maximum Likelihood Estimation Techniques," in Theory and Applications of Kalman Filtering, NATO AGARDograph 139, Feb. 1970.

5. Jazwinski, A. H., Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.

6. Sorenson, H. W., Sacks, J. E., "Recursive Fading Memory Filtering," Info. Sci., 3, 1971, pp. 101-119.

7. Wishner, R. P., Tabaczynski, J. A., Athans, M., "A Comparison of Three Non-Linear Filters," Automatica, 5, 1969, pp. 487-496.

8. Sorenson, H. W., Stubberud, A. R., "Recursive Filtering for Systems with Small but Nonnegligible Nonlinearities," Int. J. Contr. 7, 1968, pp. 271-280.

9. Sorenson, H. W., A Nonlinear Perturbation Theory for Estimation and Control of Time-Discrete Stochastic Systems, Ph.D. Dissertation, UCLA, 1966.

10. Sorenson, H. W., Stubberud, A. R., "Nonlinear Filtering by Approximation of the A Posteriori Density," Int. J. Contr., 8, 1968, pp. 33-51.

11. Bucy, R. S., "Bayes Theorem and Digital Realizations of Nonlinear Filters," J. Astr. Sci., 17, 1970, pp. 80-94.

12. Bucy, R. S., Senne, K. D., "Realization of Optimum Discrete-Time Nonlinear Estimators," Proc. 1st Symp. Nonlin. Est. Theory, San Diego, 1970, pp. 6-17.

13. Sorenson, H. W., Alspach, D. L., "Recursive Bayesian Estimation Using Gaussian Sums," Automatica, 7, 1971.

14. Alspach, D. L., Sorenson, H. W., "Approximation of Density Functions by a Sum of Gaussians for Nonlinear Estimation," Proc. 1st Symp. Nonlinear Est. Theory, San Diego, 1970, pp. 19-31.

15. Center, J. L., "Practical Nonlinear Filtering of Discrete Observations by Generalized Least-Squares Approximation of the Conditional Probability Distribution," Proc. 2nd Symp. Nonlinear Est. Theory, San Diego, 1971.

16. Tse, E., "Parallel Computation of the Conditional Mean State Estimate for Nonlinear Systems," Proc. 2nd Symp. Nonlinear Est. Theory, San Diego, 1971.

17. de Figuiredo, R. J. P., "Spline Filters," Proc. 2nd Symp. Nonlinear Est. Theory, San Diego, 1971.

18. Sorenson, H. W., "Controllability and Observability of Linear, Stochastic, Time-Discrete Control Systems," Adv. in Contr. Syst., Vol. 6, Academic Press, New York, 1968.

19. Fitts, J., "On the Observability of Nonlinear Systems with Applications to Nonlinear Regression Analysis," Proc. 1st Symp. Nonlinear Est. Theory, San Diego, 1970.

20. Griffith, E. W., Kumar, K. S. P., "On the Observability of Nonlinear Systems," J. Math. Anal. Appl., 35, 1971, pp. 135-147.

21. Alspach, D. L., "A Bayesian Approximation Technique for Estimation and Control of Time-Discrete Stochastic Systems," Ph. D. Dissertation, UCSD, 1970.

22. Kushner, H. J., "Approximations to Nonlinear Filters," IEEE Trans. Auto. Cont., AC-12, 5, Oct. 1969, pp. 546-556.

23. Athans, M., Wishner, R. P., Bertolini, A., "Suboptimal State Estimators for Continuous-Time Nonlinear Systems," IEEE Trans. Auto. Contr., AC-13, 5, Oct. 1968, pp. 504-514.

24. Schwartz, L., Stear, E. B., "A Computational Comparison of Several Nonlinear Filters," IEEE Trans. Auto. Cont., AC-13, 1, Feb. 1969, pp. 83-86.

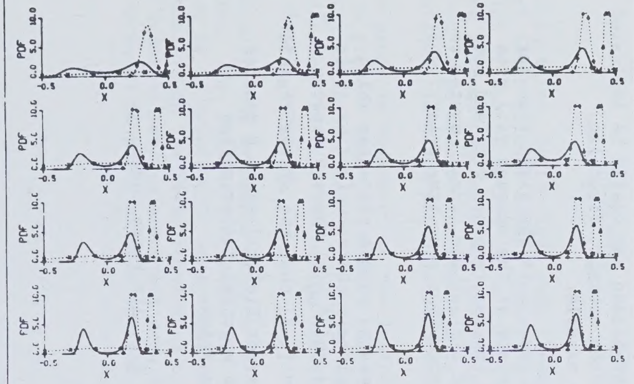


Fig. 1: No Plant Nonlinearity or Plant Noise

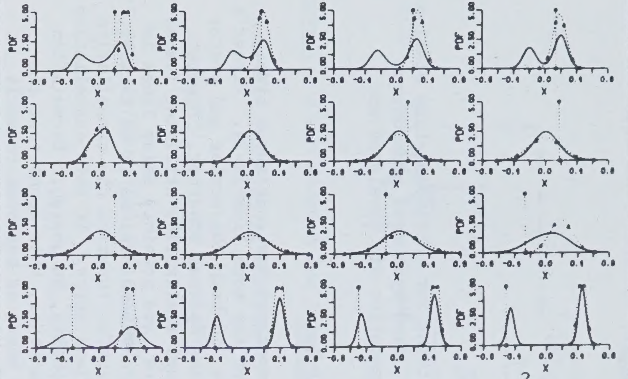


Fig. 2: No Plant Nonlinearity but $\sigma_w^2 = 0.1$

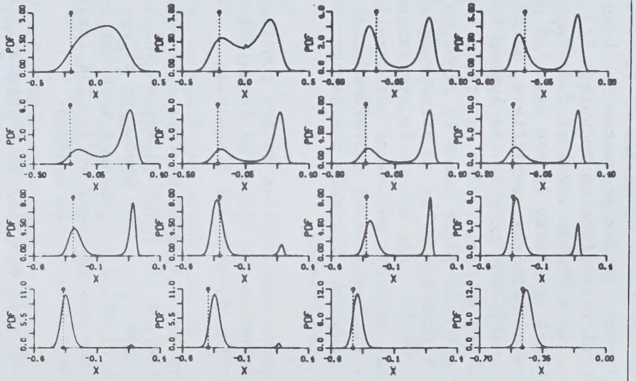


Fig. 3: Plant Nonlinearity but No Plant Noise

SYNTHESIS OF SUB-OPTIMAL FIXED LAG SMOOTHERS*

J. S. MEDITCH
University of California
Irvine, California 92664

ABSTRACT

The problem of suboptimal, fixed-lag smoothing for continuous-time stationary, random processes is examined. Two related synthesis procedures which are readily amenable to digital computation are developed and illustrated.

1. INTRODUCTION

The mathematical synthesis of fixed-lag smoothers for continuous time, stationary, random processes has been known for some time [1-3]. Wiener's spectral factorization technique [1] leads, in both the scalar and vector cases, to a stable transfer function for the optimal smoother. However, practical synthesis in the sense of physical realization utilizing lumped-parameter networks, amplifiers, attenuators, and possibly delay lines is difficult. Not only is the system infinite-dimensional as is well-known, but there generally also exists in the transfer function the cancellation of right half-plane poles by right half-plane zeros. If this cancellation is not effected prior to attempting a synthesis, an unstable formulation will result. For nonstationary random processes, this means that the optimal smoother possesses uncontrollable modes as has been recently noted and resolved by Chirarattananon and Anderson [4] for the discrete-time problem. The same has not yet been accomplished in the continuous-time case however.

The above difficulties in the continuous-time case can easily be indicated by reviewing the classical formulation of the optimal, fixed-lag, smoothing problem and discussing a simple example.

In the block diagram of Fig. 1, t denotes time, T is the fixed time-lag, and z(t) is the received signal which consists of the message y(t) plus the noise v(t). The smoothing filter has impulse response w(t) and output x(t) which is compared with the output i(t) of the "ideal" system to form the system error e(t). The ideal system is a pure time delay T which operates only on the message.

Now let y(t) and v(t) be zero-mean, stationary, random processes which are independent of each other and are completely characterized by their respective power spectral densities $\Phi_{yy}(s)$ and $\Phi_{vv}(s)$ where s denotes complex frequency. Then, it is well-known [1-3] that the system transfer function which minimizes the mean-square error $E[e^2(t)]$ where E denotes the expected value is readily obtained via spectral factorization solution of the appropriate Wiener-Hopf equation.

---

* This work was supported in part by the U.S. Air Force under Grant No. AFOSR-71-2116.

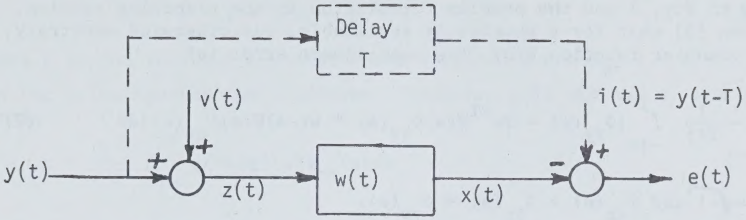


Figure 1. Block diagram for fixed-lag smoothing problem formulation.

As a specific example, if one has

$$\Phi_{yy}(s) = \frac{20}{-s^2+4} \quad \text{and} \quad \Phi_{vv}(s) = 4,$$

it is easy to show that the optimal smoother has the transfer function

$$W(s) = \frac{e^{-3T}(s+2) - 5e^{-sT}}{(s-3)(s+3)} \tag{1}$$

The system is clearly infinite dimensional. What is not quite so apparent is that the pole at s = 3 is cancelled by a zero there. However, one can verify via l'Hospital's rule that

$$\lim_{s\to 3} W(s) = \frac{(5T+1)e^{-3T}}{6} .$$

At this point, it is clear that there exist at least two problem areas which merit further study. The first is to conduct a theoretical investigation of the pole-zero cancellation phenomenon with the intent of developing general synthesis procedures for optimal, fixed-lag smoothing. The second is to examine the possibility of suboptimal smoothing synthesis techniques in which the pole-zero cancellation problem is not present.

This paper is addressed to the latter. Specifically, the purpose here is to present some preliminary results on synthesis procedures which can easily be implemented on a digital computer. The emphasis is on simplicity in physical realization of the smoothing operation with passive elements as already remarked.

Basically, the synthesis problem is reduced to one of parameter optimization. Two related procedures are then developed and illustrated. The first is an essentially analytical one, while the second which is more flexible than the first requires a constrained parameter search. Only the scalar-input/scalar-output case is treated, but the results can readily be extended to multivariable systems.

2. SUBOPTIMAL SMOOTHING

Referring to Fig. 1 and the problem formulation in the preceding section, it is known [3] that for a physically realizable, but otherwise arbitrary, smoother transfer function W(s), the mean-square error is*

$$E[e^2(t)] = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} [\Phi_{yy}(s) - 2e^{sT}W(s)\Phi_{yy}(s) + W(-s)W(s)\Phi_{zz}(s)]ds \qquad (2)$$

where $j = \sqrt{-1}$ and $\Phi_{zz}(s) = \Phi_{yy}(s) + \Phi_{vv}(s)$.

If the structure of the smoothing transfer function is constrained to within a set of parameters ξ and this is denoted by setting $W(s) = W(s;\xi)$, then it is clear that minimization of the mean-square error with respect to ξ is equivalent to minimization of

$$I = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} [W(-s;\xi)W(s;\xi)\Phi_{zz}(s) - 2e^{Ts}W(s;\xi)\Phi_{yy}(s)]ds \qquad (3)$$

In order to guarantee a stable system, this minimization must, in general, be subject to a set of nonlinear inequality constraints on ξ in order to force the poles of the minimizing $W(s;\xi)$ to lie in the left half-plane.

Assuming that the two integrals in Eq. (3) can be evaluated in parametric form (see, for example, Appendices A and E of [3]), the problem becomes one of constrained parameter optimization.

First Method

As expected, considerable simplification ensues if additional structure is imposed on $W(s;\xi)$. With this in mind, the first synthesis procedure can now be readily derived. The development is analogous to one given by Marowitz and Fegley [5] with one significant difference. Namely, the value of ξ which minimizes I is expressed analytically in the present work. Its computation then merely involves solution of a system of linear algebraic equations, or equivalently, a matrix inversion, rather than implementation of a quadratic program as in [5]. However, the quadratic programming formulation would be useful here if it were desired, for example, to impose bandwidth or rise time constraints on the smoother transfer function.

In this first procedure, $W(s;\xi)$ is taken to be of the form

$$W(s;\xi) = \sum_{k=1}^{n} \xi_k g_k(s)$$

where each $g_k(s)$, $k = 1, \cdots, n$, is a specified, stable, transfer function.

---

* The factor of 2π is included in the denominator here to avoid its persistent reappearance in certain calculations in the sequel.

In this case, Eq. (3) becomes

$$I = \xi' A \xi - 2\xi' b \tag{4}$$

where ξ is the (column) n-vector whose components are ξ_k, $k = 1, \cdots, n$, and the prime denotes the transpose. Further, A is the n X n, symmetric matrix

$$A = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} g(-s) g'(s) \Phi_{zz}(s) ds \tag{5}$$

where g(s) is the n-vector with components $g_k(s)$, $k = 1, \cdots, n$, and b is the n-vector

$$b = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} e^{Ts} g(s) \Phi_{yy}(s) ds. \tag{6}$$

It can be shown that, under fairly general restrictions on g(s) and $\Phi_{zz}(s)$, A is positive-definite.

Setting the gradient of I with respect to ξ in Eq. (4) equal to zero, it follows that

$$\xi = A^{-1} b \tag{7}$$

is both a necessary and sufficient condition for the minimization of I. The sufficiency is a consequence of the fact that A is positive-definite.

Substitution of Eq. (7) into Eq. (4) yields $I = -b'A^{-1}b$ which in view of Eq. (2) means that the corresponding mean-square error is given by

$$E[e^2(t)] = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \Phi_{yy}(s) ds - b'A^{-1}b. \tag{8}$$

Example

As a simple illustration of these results, consider the same spectra as in the example in Section 1, i.e.,

$$\Phi_{yy}(s) = \frac{20}{-s^2+4} \quad \text{and} \quad \Phi_{yy}(s) = 4$$

from which it follows that

$$\Phi_{zz}(s) = \Phi_{yy}(s) + \Phi_{vv}(s) = 4 \; \frac{-s^2+9}{-s^2+4} \; .$$

Guided by the fact that the optimal smoother has a single pole at s = -3, select the simplest possible form for the suboptimal smoother transfer function, viz.,

$$W(s;\xi) = \frac{\xi}{s+3}$$

where ξ is a scalar. In this case, A and b are also scalars, and evaluation of the integrals in Eq. (5) and (6) yields

$$A = 1 \qquad \text{and} \qquad b = 5e^{-2T} - 4e^{-3T},$$

respectively. It then follows from Eq. (7) that ξ=b so that the suboptimal smoother transfer function is

$$W(s) = \frac{b}{s+3} = \frac{5e^{-2T}-4e^{-3T}}{s+3} \tag{9}$$

For T = 0, Eq. (1) gives the transfer function

$$W_f(s) = \frac{1}{s+3}$$

for the optimal filter. The same result also follows from Eq. (9) for T = 0. In any event, it is clear that implementation of the suboptimal smoother is extremely simple, consisting merely of the gain b in cascade with the optimal filter. However, the question of whether or not this is worth doing depends upon the suboptimal smoother's performance relative to that of the filter alone. This comparison is made below. For the sake of completeness, a comparison is also made with the performance of the optimal smoother of Eq. (1).

The mean-square error for each system is readily shown to be:

1. Optimal filter: 4
2. Suboptimal smoother: $5-(5e^{-2T}-4e^{-3T})^2$
3. Optimal smoother: $\frac{2}{3}(5 + e^{-6T})$

The latter two are tabulated below for various values of the lag T along with the corresponding values of b, the suboptimal smoother gain.

It is seen that the suboptimal smoother has less mean-square error than the optimal filter only for T in the range 0 to approximately 0.45. In this range, the suboptimal smoother mean-square error achieves its minimum in the vicinity of T = 0.18 where its performance is quite comparable to that of the optimal smoother. Also at T = 0.18, the suboptimal smoother offers a reduction of about 8.5% in the mean-square error relative to the optimal filter. For $T \geqslant 0.45$, the suboptimal smoother's mean-square error is greater than that of the optimal filter and approaches the asymptotic value of 5 as T increases. The latter result is evident from the expression given above for the suboptimal smoother's mean-square error.

In performance, the suboptimal smoother is competitive with the optimal smoother from $0 \leqslant T \leqslant 0.18$. Beyond that, its performance degrades rather sharply relative to the optimal. Parenthetically, it is noted that the asymptotic value of the mean-square error for the optimal smoother is 10/3. Thus, for T > 2/3, the optimal smoother offers a reduction of about 16.7% in the mean-square error relative to the optimal filter.

T	Mean-Square Error		Gain b
	Optimal Smoother	Suboptimal Smoother	
0.0	4.0000	4.0000	1.0000
0.06	3.7984	3.8042	1.0935
0.12	3.6578	3.6948	1.1424
0.18	3.5597	3.6604	1.1574
0.24	3.4913	3.6846	1.1469
0.30	3.4435	3.7506	1.1178
0.35	3.4150	3.8267	1.0832
0.40	3.3938	3.9145	1.0419
0.45	3.3781	4.0082	.99589
0.50	3.3665	4.1034	.94688

Table 1. Optimal and suboptimal smoother performance, and suboptimal smoother gain as functions of lag T.

Second Method

The second method is a generalization of the first wherein the smoother transfer function has the assumed structure

$$W(s;\xi,\zeta) = \sum_{k=1}^{n} \xi_k g_k(s;\underline{\zeta}_k). \tag{10}$$

Here, each $g_k(s;\underline{\zeta}_k)$ is a transfer function of specified form which is parametric in a vector $\underline{\zeta}_k$, and ξ_k is the same as before.

Substitution of Eq. (10) into Eq. (3) yields

$$I = \xi'A(\zeta)\xi - 2\xi'b(\zeta) \tag{11}$$

where $\zeta = [\underline{\zeta}_1,\cdots,\underline{\zeta}_n]'$. In Eq. (11),

$$A(\zeta) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} g(-s;\zeta)g'(s;\zeta)\Phi_{zz}(s)ds \tag{12}$$

and

$$b(\zeta) = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} e^{Ts}g(s;\zeta)\Phi_{yy}(s)ds \tag{13}$$

where $g(s;\zeta) = [g_1(s;\underline{\zeta}_1),\cdots, g_n(s;\underline{\zeta}_n)]$.

Minimization of I in Eq. (11) is unconstrained with respect to ξ, but constrained with respect to some or all of the components of ζ to invoke stability. Thus setting the gradient of I with respect to ξ equal to zero yields

$$\xi = A^{-1}(\zeta)b(\zeta) \tag{14}$$

Substitution of this result into Eq. (11) gives

$$I(\zeta) = -b'(\zeta)A^{-1}(\zeta)b(\zeta) \tag{15}$$

as the function to be minimized with respect to ζ, subject to the aforementioned constraints.

This latter minimization problem is currently under investigation, but there are no concrete results to report at this time. However, it appears that the use of existing computational methods is possible. An example formulation is given below.

Example

Considering the same spectra as in the preceding example, let

$$W(s;\xi;\zeta) = \frac{\xi}{s+\zeta}$$

where ξ and ζ are scalars. The constraint here is obviously $\zeta > 0$.

In this case, Eqs. (12), (13) and (14) give

$$A(\zeta) = \frac{2\zeta + 9}{\zeta(\zeta+2)}, \qquad b = \frac{5[e^{-2T}(\zeta+2)-4e^{-\zeta T}]}{\zeta^2-4},$$

and

$$\xi = \frac{5\zeta[e^{-2T}(\zeta+2)-4e^{-\zeta T}]}{(2\zeta+9)(\zeta-2)}.$$

From Eq. (15) one obtains

$$I(\zeta) = -\frac{25\zeta[e^{-2T}(\zeta+2)-4e^{-\zeta T}]^2}{(2\zeta+9)(\zeta-2)^2(\zeta+2)}$$

as the function to be minimized subject to $\zeta > 0$.

3. CONCLUSION

Of the two synthesis methods presented, the first is clearly more attractive computationally. However, it requires that pole-zero locations be specified a priori. While the optimal smoother solution can be used as a guide in this case, there is no assurance that this procedure is "best" in any sense other than perhaps expediency. It appears from preliminary work that systematic procedures for selecting the pole-zero locations for each $g_k(s)$, k=1,•••,n, are possible using orthonormal functions and their corresponding network realizations. If successful, this work should make the first method extremely attractive.

The second method could also be enhanced in a slightly different way. Namely, if the search over ζ could be justifiably constrained to a fixed grid in the complex plane, the amount of computation required would be reduced significantly.

Finally, in both methods it is necessary to calculate the mean-square error explicitly in order to compare suboptimal with optimal performance. A means for determining error bounds would clearly expedite this process.

REFERENCES

1. N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," Technology Press, Cambridge, Mass., 1949.

2. Y.W. Lee, "Statistical Theory of Communication," John Wiley and Sons, New York, 1960.

3. G.C. Newton, Jr., L.A. Gould, and J.F. Kaiser, "Analytical Design of Linear Feedback Controls," John Wiley and Sons, New York, 1957.

4. S. Chirarattananon and B.D.O. Anderson, "The Fixed-Lag Smoother as a Stable Finite-Dimensional Linear System" to appear.

5. J.E. Marowitz and K.A. Fegley, "System Design Using Quadratic Programming," IEEE Trans. Auto. Cont., Vol. AC-16, No. 3, June 1971, pp. 241-247.

PARAMETER ESTIMATION IN PARTIALLY SPECIFIED STOCHASTIC SYSTEMS†

R. L. KASHYAP
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

ABSTRACT

A partially specified stochastic linear system is one in which the input disturbance is uncorrelated with only a finite number of the previous values of the noise, output and the control input. The remaining higher order correlation coefficients of the noise are unspecified. In particular, the input disturbances need not be a discrete white noise. We will develop a recursive algorithm for the estimation of the coefficients in a partially specified linear system and prove the consistency of the estimates. Simulation studies indicate that the variance of the estimate is of the same order as the Cramer-Rao lower bound, even for sample sizes as small as 50.

I. INTRODUCTION AND SUMMARY

There have been a number of papers on the identification of stochastic linear dynamical systems [1-10]. In all of these papers, one assumes that the process $y(\cdot)$ under consideration obeys a linear system (1.1) excited by the control input $v(\cdot)$ which can be measured exactly at every instant and the disturbance input $\xi(\cdot)$ which cannot be observed directly, obeying (1.2).

$$y(i) = \Sigma_{j=1}^{n_4} \alpha_j y(i-j) + \Sigma_{j=1}^{n_5} \beta_j v(i-j) + \xi(i) + \Sigma_{j=1}^{n_6} \gamma_j \xi(i-j) \qquad (1.1)$$

$$\left.\begin{array}{l} E[\xi(k)] = 0, \quad E[\xi(k)\xi(j)] = R_\xi \delta_{kj} \quad , \\ E[\xi(k)y(k-j)] = 0, \quad E[\xi(k)v(k-j)] = 0, \quad j \geq 1 \end{array}\right\} \qquad (1.2)$$

α_j, β_j, γ_j are the undetermined parameters. The restrictive assumptions (1.2) on the noise $\xi(k)$ are introduced mainly so that the parameters α_j, ets., can be estimated with ease using the available measurements $\{y(j), j=1,\dots,N\}$. However, while dealing with stochastic processes occuring in nature such as rainfall, riverflow, etc., assumptions such as (1.1) and (1.2) appear to be highly unrealistic unless the integers n_4, n_5, and n_6 are very large. Consequently, we consider models involving assumptions far less restrictive than (1.1) and (1.2). Our basic model is described in (1.3) where the corresponding noise process $w(k)$ obeys (1.4).

$$y(i) = \alpha_o + \Sigma_{j=1}^{n_1} \alpha_j y(i-j) + \Sigma_{j=1}^{n_2} \beta_j v(i-j) + w(i) + \Sigma_{j=1}^{n_3} \gamma_j w(i-j) \qquad (1.3)$$

$$\left.\begin{array}{ll} E[w(i)y(i-j)] = 0, & j=1,\dots,n_1 \\ E[w(i)v(i-j)] = 0, & j=1,\dots,n_2 \\ E[w(i)w(i-j)] = 0, & j=1,\dots,n_3 \end{array}\right\} \qquad (1.4)$$

The integers n_1, n_2 and n_3 in (1.3) can be chosen by us, and the higher order correlations such as $E[w(i)\ y(i-j)]$, $j > n_1$, $E[w(i)w(i-j)]$, $j > n_3$, are unspecified. Hence we call the model (1.3)-(1.4) a partially specified model. We can show that if the process $y(\cdot)$ and $v(\cdot)$ are stationary, then

---

†This work was partially supported by NSF under grant GK-18225.

the representation (1.3)-(1.4) is valid for any n_1, n_2, and n_3. Moreover, if the processes $y(\cdot)$ and $v(\cdot)$ do indeed obey the model (1.1)-(1.2) with n_4, n_5, and n_6 known, and if we choose $n_1 = n_4$, $n_2 = n_5, n_3 = n_6$, then the coefficients α_j, β_j, γ_j associated with (1.3) are identical to the corresponding coefficients in (1.1). Thus, (1.1)-(1.2) is a very special case of the model in (1.3)-(1.4).

We develop a recursive algorithm to estimate α_j, β_j, and γ_j. Let $a_j(N)$ be the corresponding estimate of α_j based on the measurements $\{y(j), v(j), j \leq N\}$ (or $\{y(j), j \leq N\}$). We define $b_j(N)$ and $c_j(n)$ in a similar manner. We will show that the estimates $a_j(N)$, $b_j(N)$ and $c_j(N)$ tend to their corresponding values α_j, β_j, and γ_j as N tends to infinity. We will illustrate the properties of the estimate by means of a numerical example.

II. THE BASIC MODEL

Let us introduce the backward shift operator z defined by the relation $zy(t) = y(t-1)$. z will also be used as a complex variable $z = \exp(-i\lambda)$ where λ is the complex frequency variable. Let

$$A(z) = 1 - \Sigma_{j=1}^{n_1} \alpha_j z^j; \quad B(z) = \Sigma_{j=1}^{n_2} z^j; \quad C(z) = 1 + \Sigma_{j=1}^{n_3} \gamma_j z^j \tag{2.1}$$

Then the model in (1.3) can be rewritten as in (2.2).

$$A(z)y(i) = \alpha_o + B(z)v(i) + C(z)\xi(i) \tag{2.2}$$

We make the following assumptions (A1)-(A3).

(A1) All the zeros of $C(z)$ lie in the region $|z| > 1$.

(A2) All the zeros of $A(z)$ and $B(z)$ should be in the region $|z| > 1$.

(A3) The polynomials A, B, and C do not have zeros which are common to all of them.

The assumptions (A1)-(A2) ensure that the $w(\cdot)$, $y(\cdot)$ and $v(\cdot)$ sequences are stationary. (A3) ensures that the model (2.2) is completely controllable and observable.

The feasibility of the representation of the processes $y(\cdot)$ and $v(\cdot)$ in the form of Equation (1.3)-(1.4) is stated in proposition 1.

Proposition 1. Let $y(\cdot)$ and $v(\cdot)$ be stationary stochastic processes. Let the process $w(\cdot)$ be recursively generated from Equation (1.3). Then the coefficients α_j, β_j, and γ_j can be chosen so as to satisfy Equation (1.4). Moreover, the corresponding polynomials $A(z)$, $B(z)$ and $C(z)$ will be unique if we impose the restrictions (A1)-(A3).

The proof of the proposition is skipped for lack of space.

A special case of the model (1.3)-(1.4) is when $w(\cdot)$ is a discrete white noise. A more general case is when the noise in (1.3) is white but $y(k)$ and $v(k)$ cannot be observed exactly. Let $y'(k)$ and $v'(k)$ be the corresponding observed processes and let them be related to $y(k)$ and $v(k)$ through additive noises. Then one can show [16] that $y'(k)$ and $v'(k)$ obey an equation of form (1.3) in which the additive noise, say $w'(k)$, is white.

III. ALGORITHM FOR THE ESTIMATION OF THE COEFFICIENTS

3.1 Development of the Algorithm

We will relabel the vector of undetermined parameters in (1.3)-(1.4) as $\underset{\sim}{\alpha}$ where

$$\underset{\sim}{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{n_1}, \beta_1, \ldots, \beta_{n_2}, \gamma_1, \ldots, \gamma_{n_3})^T, \quad n = 1 + n_1 + n_2 + n_3$$

We would like to estimate the parameter α from the sample values of y(·) and v(·). Let a(k) be an estimate of α based on the measurements {y(j), v(j), $j \leq k$}.

Let $\hat{y}(k+1|k)$ = predicted value of y(k+1) given all measurements y(j), v(j), $j \leq k$ and the coefficient α. Equation (1.3) suggests the following recursive formula for $\hat{y}(k+1|k)$.

$$\hat{y}(k|k-1) = \alpha_0 + \Sigma_{j=1}^{n_1} \alpha_j y(k-j) + \Sigma_{j=1}^{n_2} \alpha_{n_1+j} v(k-j) +$$

$$\Sigma_{j=1}^{n_3} \alpha_{n_1+n_2+j} [y(k-j)-\hat{y}(k-j|k-j-1)]. \qquad (3.1)$$

If the coefficients α_j are unknown, the corresponding one-step predictor of y(k) will be called $\bar{y}(k|k-1)$ whose expression, as suggested by (3.1) is given in (3.2), where a is an estimate of α and $\bar{w}(j)$ is the available estimate of w(j) at time instant j.

$$\bar{y}(k|k-1) = \utilde{a}^T \utilde{z}(k-1) \qquad \utilde{a}^T = (a_0, a_1, \ldots, a_{n-1}) \qquad (3.2)$$

$$\utilde{z}^T(k-1) = [1, y(k-1), .., y(k-n_1), v(k-1), .., v(k-n_2), \bar{w}(k-1), .., \bar{w}(k-n_3)]. \quad (3.3)$$

Assuming that we have the estimates $\bar{w}(j)$, $j \leq k$, we will compute a(k+1) by minimizing the following criterion function $\bar{J}_{k+1}(a)$, which is nothing but the sum of squares of the one-step prediction errors until the time (k+1).

$$J_{k+1}(\utilde{a}) = \Sigma_{j=1}^{k+1} (y(j)-\bar{y}(j|j-1))^2 = \Sigma_{j=1}^{k+1} (y(j)-\utilde{a}^T \utilde{z}(j-1))^2 \qquad (3.4)$$

Performing the indicated minimization, we get

$$\partial J_{k+1}(\utilde{a})/\partial \utilde{a} = 0 \Rightarrow \Sigma_{j=1}^{k+1} (y(j)-\utilde{a}^T \utilde{z}(j-1)\utilde{z}(j-1)) = 0 \qquad (3.5)$$

Let $\utilde{a}(k+1)$ be the minimizing value of a. Then from (3.5) we get

$$\utilde{a}(k+1) = \utilde{S}(k+1)\utilde{d}(k+1) \qquad (3.6)$$

where

$$\utilde{S}(k+1) = [\Sigma_{j=1}^{k+1} \utilde{z}(j-1)\utilde{z}^T(j-1)]^{-1}, \ \utilde{d}(k+1) = \Sigma_{j=1}^{k+1} y(j)\utilde{z}(j-1). \qquad (3.7)$$

Using $\utilde{a}(k+1)$, we obtain the following estimate $\bar{w}(k+1)$ of w(k+1).

$$\bar{w}(k+1) = y(k+1) - \utilde{a}^T(k+1)\utilde{z}(k) \quad . \qquad (3.8)$$

Since we have obtained $\bar{w}(k+1)$, we can construct the vector z(k+1) and the estimate a(k+2) from (3.6) by replacing (k+1) by (k+2). This process can be repeated.

The formula for a(k) and S(k) can be rewritten recursively so as to avoid explicit inversion of a matrix at every instant. The complete algorithm is given below, where we have rewritten (3.8) and (3.3).

$$\left.\begin{aligned}
&\utilde{S}(k+1) = \utilde{S}(k)-\utilde{S}(k)\utilde{z}(k)\utilde{z}^T(k)\utilde{S}(k)/(1+\utilde{z}^T(k)\utilde{S}(k)\utilde{z}(k)) \\
&\utilde{a}(k+1) = \utilde{a}(k)+\utilde{S}(k+1)\utilde{z}(k)(y(k+1)-\utilde{a}^T(k)\utilde{z}(k)) \\
&\bar{w}(k+1) = y(k+1)-\utilde{a}^T(k+1)\utilde{z}(k) \\
&\utilde{z}(k) = (1, y(k), .., y(k-n_1+1), v(k), .., v(k-n_2+1), \ \bar{w}(k), .., \bar{w}(k-n_3+1))^T
\end{aligned}\right\} \ (C1)$$

Initialization: If the integer $n_3 = 0$, then the algorithm is not very sensitive to the starting value, and we can use the algorithm (C1) for $k \geq r$

with $\underset{\sim}{a}(r) = \underset{\sim}{0}$ and $\underset{\sim}{S}(r) = I$ where r is any integer greater than n .

If $n_3 > 0$, the numerical convergence of the algorithm seems to be very dependent on the initial choice of $\underset{\sim}{a}(r)$ and $\underset{\sim}{S}(r)$, especially if the moving average terms in (1.3) predominate. In such cases, the following choice is very helpful.

$$\underset{\sim}{S}(r) = \left[\Sigma_{j=n_4+1}^{r} \underset{\sim}{z}(j-1)\underset{\sim}{z}^T(j-1)\right]^{-1}, \quad \underset{\sim}{d}(r) = \Sigma_{j=1}^{r} y(r)\underset{\sim}{z}(j-1), \quad \underset{\sim}{a}(r)=\underset{\sim}{S}(r)\underset{\sim}{d}(r)$$

The vectors $z(n_4),\ldots,z(r-1)$ are constructed as indicated in (C1) using the observed measurements $\{y(j),v(j), j \leq r\}$ and the numbers $\bar{w}(j)$, $j=1,\ldots,r$ which are taken from a table of Gaussian random numbers with zero mean and variance equal to the sample variance of $y(j)$. The algorithm (C1) is used to construct $\underset{\sim}{a}(j)$, $j \geq r + 1$. The integer r is very much greater than n .

The algorithm was first suggested in [7]. However, its use in partially specified systems and the associated proposition 2 were not mentioned there.

3.2 Convergence Properties of the Algorithm

Proposition 2. We make the following assumptions:

(A4) The processes $y(\cdot)$ and $v(\cdot)$ are ergodic obeying (1.3) and (1.4). The coefficient vector α obeying (A1)-(A3).

(A5) The sequence $a(k)$ given by algorithm (C1) converges to a value $\bar{a}$ with probability one as k tends to infinity.

If (A 4) and (A 5) are satisfied, then $\underset{\sim}{\bar{a}} = \underset{\sim}{\alpha}$

Proof: Construct the residuals $w(k)$ from (3.10).

$$\begin{aligned} w(k) &= y(k) - \bar{a}_o - \Sigma_{j=1}^{n_1} \bar{a}_j\, y(k-j) - \Sigma_{j=1}^{n_2} \bar{a}_{n_1+j}\, v(k-j) \\ &\quad - \Sigma_{j=1}^{n_3} \bar{a}_{n_1+n_2+j}\, w(k-j) \ . \ . \ . \qquad (3.10) \\ &= y(k) - \underset{\sim}{\bar{a}}^T\, \underset{\sim}{z}(k-1) \end{aligned}$$

By (A4) and (A5), the process $w(\cdot)$ is ergodic and consequently we have (3.11)

$$\begin{aligned} &\lim_{k \to \infty} k\underset{\sim}{S}(k) = \underset{\sim}{S}_\infty \triangleq \{E[\underset{\sim}{z}(j-1)\underset{\sim}{z}^T(j-1)]\}^{-1} \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{W.P.1} \qquad\qquad (3.11) \\ &\lim_{k \to \infty} \underset{\sim}{d}(k)/k = \underset{\sim}{d}_\infty \triangleq E[y(j)\underset{\sim}{z}(j-1)] \end{aligned}$$

By the definition of $\underset{\sim}{\bar{a}}$,

$$\underset{\sim}{S}_\infty^{-1}\, \underset{\sim}{\bar{a}} = \underset{\sim}{d}_\infty \qquad (3.12)$$

Combining (3.11) and (3.12), we get

$$\begin{aligned} &E[\underset{\sim}{z}(j-1)\ \underset{\sim}{z}^T(j-1)]\ \underset{\sim}{\bar{a}} = E[y(j)\underset{\sim}{z}(j-1)] \qquad \text{or} \\ &E[\underset{\sim}{z}(j-1)\ w(j)] = 0 \qquad\qquad (3.13) \end{aligned}$$

Utilizing the definition of $\underset{\sim}{z}(\cdot)$ in (C1), we get from (3.13) the following equations

$$\left.\begin{aligned} &E[w(k)] = 0 \quad , \quad E[w(k)y(k-j)] = 0, \quad j=1,\ldots,n_1 \\ &E[w(k)w(k-j)] = 0, j=1,\ldots,n_2; \quad E[w(k)w(k-j)] = 0, \quad j=1,\ldots,n_3 \end{aligned}\right\} \ (3.13)$$

Let us denote $\bar{\bar{b}}_j = \bar{a}_{n_1+j}$, $j=1,\ldots,n_2$ and $\bar{c}_j=a_{n_1+n_2+j}$, $j=1,\ldots, n_3$. Then the polynomials $\bar{A}(z)$ and $\bar{B}(z)$ and $\bar{C}(z)$ constructed from the coefficient vector $\bar{\underset{\sim}{a}}$, as in (2.1), satisfy the conditions (A1)-(A2) in view of the ergodicity of the processes $y(\cdot)$, $v(\cdot)$ and $w(\cdot)$. This fact coupled with (3.14) and proposition 1 implies that $\bar{\underset{\sim}{a}}$ is independent of the starting value of $\underset{\sim}{a}(o)$ and $\bar{\underset{\sim}{a}} = \underset{\sim}{\alpha}$. (Q.E.D.)

To place the method in the traditional setting, we will make the assumption (A6) that the process $y(\cdot)$ is generated by a linear system excited by white noise.

(A6): $y(k)$ and $v(k)$ obey the following difference equation where n_4, n_5, and n_6 are finite and noise $\xi(k)$ obeys (1.2).

$$A^*(z)y(k) = \alpha_0^* + B^*(z)v(k) + C^*(z)\xi(k), \quad \alpha^*_{n_4} \beta^*_{n_5} V^*_{n_6} \neq 0$$

$$A^*(z) = 1 - \Sigma_{j=1}^{n_4} \alpha_j^* z^j, \quad B^*(z) = \Sigma_{j=1}^{n_5} \beta_j^* z^j; \; C^*(z) - 1 + \Sigma_{j=1}^{n_6} \gamma_j^* z^j$$

The polynomials A^*, B^*, and C^* obey the assumptions (A1)-(A3). The spectrum $S_{vv}(z)$ of $v(k)$ is arbitrary.

Since n_4, n_5, and n_6 are, in general, not known, we will consider the possibility of overestimating and underestimating them.

Case (i). We underestimate one or all of the integers n_4, n_5, n_6; i.e., either $n_1 < n_4$ or $n_2 < n_5$ or $n_3 < n_6$. Here the parameters α, β, γ characterizing the system (2.1)-(2.2) are unique and $\bar{a} = \alpha$, $\bar{b} = \beta$, $\bar{c} = \gamma$. We may have to stress the fact that, in general, $\bar{a}_j = \alpha_j \neq \alpha_j^*$, $\bar{b}_j = \beta_j \neq \beta_j^*$, etc.

Case (ii). The integers n_4, n_5, n_6 are precisely known. Here $\bar{a}_j = \alpha_j = \alpha_j^*$, $\bar{b}_j = \beta_j = \beta_j^*$, $\bar{c}_j = \gamma_j^*$.

Case (iii). All the integers n_4, n_5, and n_6 are overestimated. Then

$$\begin{aligned} \bar{a}_i &= \alpha_i^*, && i=1,\ldots,n_4 \\ &= 0, && i=(n_4+1),\ldots,n_1 \\ \bar{b}_i &= \beta_i^*, && i=1,\ldots,n_5 \\ &= 0, && i=(n_5+1),\ldots,n_2 \\ \bar{c}_i &= c_i^*, && i=1,\ldots,n_6 \\ &= 0, && i=(n_6+1),\ldots,n_3 \end{aligned}$$

We also observe that cases (ii) and (iii) rarely occur when we deal with the stochastic processes which occur in nature.

IV. NUMERICAL EXAMPLE

We will generate the $y(\cdot)$ sequence by the following difference equation:

$$y(k) = \alpha_0 y(k-1) + \xi(k) + \beta_0 \xi(k-1), \quad \alpha_0 = 0.5, \quad \beta_0 = 0.5 \quad . \qquad (4.1)$$

The sequence $\xi(k)$ is independent and has a Gauss distribution N(0,1). We will pretend that the detailed mechanism of the $y(\cdot)$ process is unknown and construct the following three models for the $y(i)$ process.

Model I. $y(k) = \alpha_1 y(k-1) + w_1(k) + \beta_1 w_1(k-1)$ where α_1, β_1 are chosen so that $E[w_1(k)y(k-1)] = 0$ and $E[w_1(k)w_1(k-1)] = 0$. It is clear that $\alpha_1 = 0.5$, $\beta_1 = 0.5$, $R_1 = E[w_1^2(k)] = 1$.

Model II. $y(k) = \alpha_2 y(k-1) + w_2(k)$ where α_2 is chosen so that $E[w_2(k)y(k-1)] = 0$. It is obvious that the true value of α_2 does not equal α_0. $\alpha_2 = E[y(k)y(k-1)] / E[y^2(k-1)] = 0.714$, $R_2 = E[w_2^2(k)] = 1.142$.

Model III. $y(k) = w_3(k) + \beta_3 w_3(k-1)$ where β_3 is chosen so that $E[w_3(k)w_3(k-1)] = 0$. $\beta_3 = 0.5[\beta_5-(\beta_5^2 - 4 - 4\alpha_o \beta_5)^{1/2}] = 0.88$, $\beta_5 = E(y^2(k)/\beta_o)$ $R_3 = E[w_3^2(k)] = E[y^2(k)]/(1 + \beta_3^2) = 1.312$

One hundred independent sets of $y(k)$, $k=1,\dots,500$ were generated and they are used to estimate the various parameters in models I, II, and III, using algorithm (C1).

Mean and Variance of Estimates. For each parameter, we compute its estimate based on the measurements $\{y(k), k=1,\dots,N\}$ for $N=50, 100,., 500$. In each case, we find the mean and variance of the 100 sets of estimates of each parameter obtained from the 100 different runs. The means are listed in Table 1 and the variances are plotted in Figure 1. It is clear that the means of the estimates are very close to their true values even when the number of observations used is as small as 50. The variance of the estimate based on N observations is proportional to $(1/N)$. In the Figure 1, we have also plotted the Cramer-Rao lower bound [13,14] on the variance of the estimates of the various parameters α_1, and β_1 . In general, the variance of the estimates given by our method is roughly 1.2 - 2.5 times that of the corresponding C-R lower bound.

Table 1: THE MEAN OF THE ESTIMATES OF PARAMETERS
IN MODELS I-III BASED ON 100 SETS OF DATA

			The mean value of the estimate based on N samples			
Model	Parameter	True Value	N=50	N=100	N=200	N=300
I	α_1	0.5	.445	.477	.496	.500
	β_1	0.5	.539	.515	.499	.498
	R_1	1.0	.909	.944	.962	.975
II	α_2	.714	.686	.701	.711	.714
	R_2	1.142	1.09	1.108	1.112	1.124
III	β_3	.88	.873	.865	.871	.881
	R_3	1.312	1.237	1.294	1.302	1.305

The recovery of integers n_1, n_2 and n_3: we also ran the algorithm with models IV, V and VI which have more unknown parameters than necessary.

Model IV. $y(k) = \alpha_4 y(k-1) + \alpha_5 y(k-2) + w_4(k) + \beta_4 w_4(k-1)$.

Model V. $y(k) + \alpha_6 y(k-1) + w_6(k) + \beta_5 w_5(k-1) + \beta_6 w_5(k-2)$.

Model VI. $y(k) = \alpha_7 y(k-1) + \alpha_8 y(k-2) + w_6(k) + \beta_7 w_6(k-1) + \beta_8 w_6(k-2)$.

Using the same set of N measurements we compute the average mean square prediction error $R(N)$.

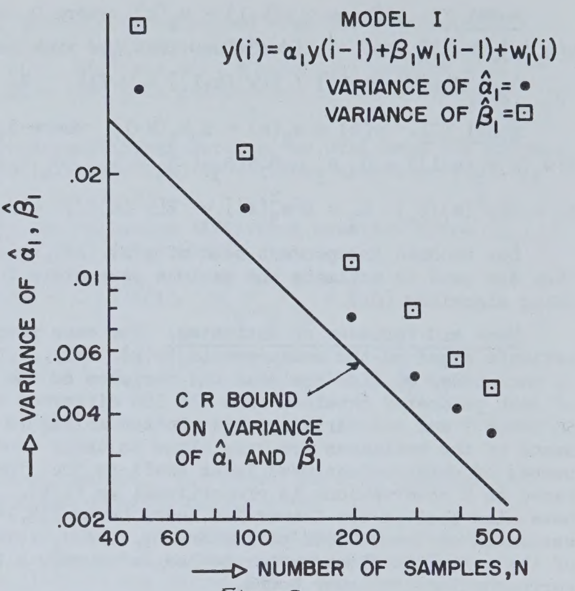


Figure 1.

$$R(N) = (1/N)\ \Sigma_{i=1}^{N}\ e^2(i)$$

$$e(i) = y(i) - \underset{\sim}{\alpha}^T(i-1)\underset{\sim}{z}(i-1)$$

= strict prediction error obtained in predicting $y(i)$ using only $y(j)$, $j < i$

It was found that in the majority of the 100 simulations, the model I gave the least value of R(N), N being as small as 100. In other words, the underestimation or overestimation of the orders of the difference equation usually results in an increase in the mean square prediction error. The prediction errors of the models with the number of parameters greater than 4 were very much larger than these of Model I in every one of the 100 runs.

V. DISCUSSION AND CONCLUSION

One of the principal aspects of our paper is the use of the partially specified model (1.3) - (1.4). The validity of the scheme for the estimation of parameters does not require the whiteness of the input noise. This property is not valid for most of the parameter estimation schemes given in the literature [2-10].

If we are dealing with the system in (1.1)-(1.2) with n_4, n_5, and n_6 known, there are a number of methods for the estimation of the parameters such as maximum likelihood (ML) [2,3], the nonlinear iterative least squares methods (NILES) [10], method of moments [11], stochastic approximation (SA) [9], and the model adjustment procedure [5]. Among these, the variance of the estimates of the parameters appears to be the least in ML and NILES methods and this variance is of the same order of the variance of the estimates in our method. The variance of the estimates in the SA and model adjustment procedures are usually very large. The ML, NILES, and the method of moments are not recursive. At every instant, we have to store the entire measurement history. There is no easy way of updating the estimates as additional measurements are obtained. Thus, it is computationally prohibitive to investigate such questions as the variability of the estimates with sample size by these methods.

Finally we come to the question of the choice of the order of the difference equation for the model. One suggestion is to assume the validity of (1.1)-(1.2) and determine the order of the difference equation by the method of hypothesis testing and chi squared tests [11,12]. There are two main objections to this procedure. First of all, the methodology involved, i.e., the posing of the problem of choice as one hypothesis testing, is open to question since the validity of the Equations (1.1)-(1.2) is usually not known in advance. Secondly, the method as developed in Ref. [11], pp. 290-293, is open to question on account of the arbitrary manner in which the number of parameters is defined. In contrast, we view the choice of the order of the difference equation as an approximation problem, and our criterion for approximation is the minimum one-step prediction error. Our method seems to bring out the effect of choosing the order of the difference equation higher than the necessary number. This aspect has also been observed in our studies on naturally occuring stochastic processes such as riverflows.

REFERENCES

1. H. B. Mann and A. Wald (1943), "On the Statistical Treatment of Linear Stochastic Difference equations", Econometrica, Vol. 11, pp. 173-220.
2. K. J. Astrom, T. Bohlin and S. Wensmark (1965), "Automatic Construction of Linear Stochastic Dynamic Models", TP 18.150, IBM Nordic Lab., Sweden, June.
3. R. L. Kashyap (1970), "Maximum Likelihood Identification of Stochastic Linear Systems", IEEE Trans. on Automatic Control, Vol. AC-15, Feb., pp. 25-34.
4. Y. C. Ho and R. C. K. Lee (1965), "Identification of Linear Dynamic Systems", J. Information and Control, Vol. 8, Feb., pp. 93-110.
5. K. Y. Wong and E. Polak (1967), "Identification of Linear Discrete Time Systems Using the Instrumental Variable Methods", IEEE Trans. on Automatic Control, Vol. AC-12, Dec., pp. 707-718.
6. H. Akaike (1970), "On a Decision Procedure of System Identification", Preprints of papers for IFAC Kyoto Symposium on Systems Engineering Approach to Computer Control, Kyoto, Japan, Aug., pp. 485-490.
7. V. Panuska (1969), "An Adaptive Recursive, Least Squares Identification Algorithm", Proc. 1969 IEEE Symposium on Adaptive Processes (8th), Paper 6-3.
8. W. N. Anderson, et.al. (1969), "Consistent Estimates of Parameters of a Linear System", Ann. Math. Stat., Vol. 40, Dec., pp. 2064-2075.
9. D. J. Sakrison (1967) , "The Use of Stochastic Approximation to Solve Identification Problem", IEEE Trans. on Automatic Control, Vol. AC-12, Oct. pp. 563-567.
10. H. Wold (1966), "Nonlinear Estimation by Iterative Least Squares Procedures", in Research Papers in Statistics, (Festschrift for J. Neyman) (Ed.) F. N. David, Wiley, pp. 411-444.
11. G. E. P. Box and G. M. Jenkins (1970), Time Series Analysis, Holden-Day.
12. J. F. Hannan (1960), Time Series Analysis, Methuen, Ch. IV, pp. 42,44.
13. K. J. Astrom (1969), "On the Achievable Accuracy in Identification Problems", IV Congress of IFAC, Warsaw.
14. C. Radhakrishna Rao (1965), Linear Statistical Inference and Its Applications, Wiley, pp. 265-266.
15. J. L. Doob (1954), Stochastic Processes, Wiley, pp. 464-466.
16. R. L. Kashyap (1970), "A New Method for Recursive Estimation in Discrete Linear Systems", IEEE Trans. on Automatic Control, Vol. AC-15, Feb., pp. 18-24.

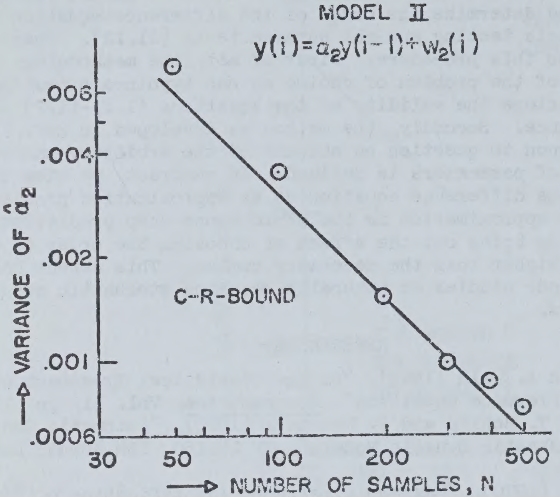


FIGURE 2

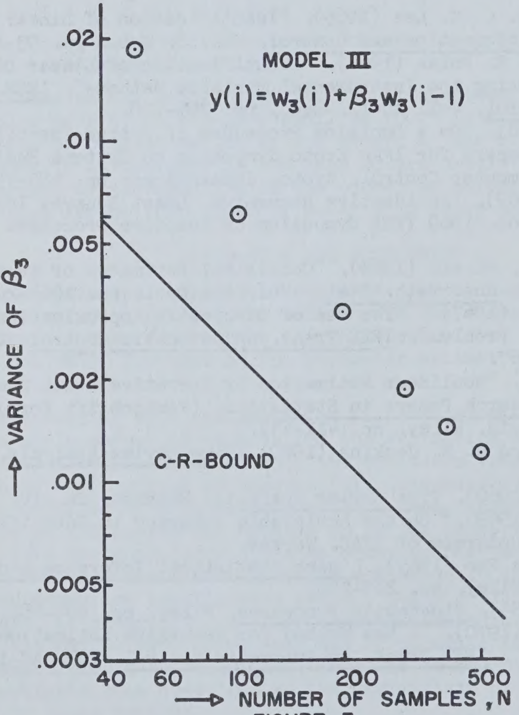


FIGURE 3

PARAMETER DISCONTINUITIES IN OPTIMALLY CONTROLLED STOCHASTIC DISCRETE SYSTEMS

P.N. NIKIFORUK, M.M. GUPTA, AND L. WAGNER
Division of Control Engineering
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

ABSTRACT

In this paper our preliminary work on the analysis and stochastic characterization of optimally controlled discrete linear systems in the presence of environmental noise (additive input and measurement noise) and subject to jump type parameter discontinuity is presented.

The procedure of analysis is to solve the state evolution process, following the parameter discontinuity by setting the new initial conditions on the process evolution process, the new initial conditions being dependent upon the process state just prior to the parameter discontinuity. The procedure is outlined for a general class of discrete systems.

The problem is treated in the stochastic sense in that an ensemble of processes with all possible initial conditions at the instant of the parameter discontinuity is considered. For general linear discrete systems, stochastic behavior in the system state, filter state, control strategy and performance index is characterized in terms of covariance matrices. Some preliminary experimental results are also given.

PROBLEM FORMULATION

Given: The process model Σ_p is defined by the following relations:

(i) state evolution process of the controlled plant:

$$\Sigma_p:\quad x(k+1) = f[x(k),\ u(k),\ w(k),\ \alpha_0 + \Delta\mu_\alpha(k-\lambda),\ k] \qquad k \varepsilon\ [k_0, k_f] \tag{1}$$

with initial conditions

$$x(k_0) = x_0 \tag{2}$$

(ii) m-dimensional nonlinear measurements:

$$y(k) = h[x(k),\ v(k),\ k] \tag{3}$$

and (iii) the performance measure, which is a real-valued scalar functional, is measured in the expected sense:

$$J = E\left\{ \sum_{k=k_0}^{k_f} L\ [x(k),\ u(k),\ k]\right\} \tag{4}$$

In (1-3), f[.] is an n-dimensional vector valued function of the plant state $x(k)\varepsilon\ R_n$, control signal $u(k)\varepsilon\ R_u$, input disturbance vector $w(k)\varepsilon\ R_w$, nominal parameter vector $\alpha_0 \varepsilon\ R_p$, and parameter discontinuity uncertainty $\Delta\alpha\mu_\alpha(k-\lambda)\varepsilon\ R_p$, where $\Delta\alpha$ is the amplitude of the discontinuity on α_0, and

$\mu_\alpha(k-\alpha)$ is a unit step function operating on α at $k=\lambda$. The measurements $y(k) \varepsilon R_m$ are corrupted by the measurement noise $v(k) \varepsilon R_v$. The disturbances $w(t)$ and $v(t)$ form a set called environmental disturbance.

Problem: For the process Σ_p, as described above, statistically characterize the dynamic effects of the parameter discontinuity $\Delta\alpha\mu_\alpha(k-\lambda)$ on the overall estimation and control problem.

METHOD OF SOLUTION

As discussed above, the state evolution process Σ_p starts at time $k=k_0$. with initial conditions $x(k_0) = x_0$, and the parameter discontinuity occurs at $k=\lambda \; \varepsilon [k_0,k_f]$. Let the time-state-space be divided into two subspaces s_1 and s_2 by a hyper-curve $g(x,\lambda)=0$, such that

$$x(k) \; \varepsilon s_1 \text{ if } g(x,k) < 0, \quad k \; \varepsilon \; [k_0,\lambda] \tag{5a}$$

and $$x_p(k) \; \varepsilon s_2 \text{ if } g(x,k) > 0, \quad k \; \varepsilon \; [\lambda,k_f] \tag{5b}$$

where, at the stage of discontinuity, the following state continuity equation must be satisfied.

$$x(k)\big|_{k=\lambda} = x_p(k)\big|_{k=\lambda} \quad \text{at } g(x,\lambda) = 0 \tag{6}$$

Thus, the state evolution processes of the plant with and without parameter uncertainty are given by

$$\Sigma: \quad x(k+1) = f\,[x(k),\, u(k),\, w(k),\, \alpha_0,\, k],\; x_0, \quad \text{for } x(k) \; \varepsilon s_1, \quad k \; \varepsilon \; [k_0,\, \lambda] \tag{7a}$$

and $$\Sigma_p: \quad x_p(x+1) = f_p\,[x_p(k),\, u_p(k),\, w(k),\, \alpha_p(k),\, k],\; x_p(\lambda) \quad \text{for } x_p(k) \; \varepsilon s_2, \quad k \; \varepsilon \; [\lambda,k_f] \tag{7b}$$

where $\alpha_p(k)$ is the perturbed parameter given by

$$\alpha_p(k) = \alpha_0 + \Delta\alpha\mu_\alpha(k-\lambda) \tag{8}$$

Comments: This method of solution is most general, and can be applied to a large class of optimally controlled processes with discontinuous type parameter uncertainties. However, since the combined problem of estimation and control for such a wide class of processes is not yet fully developed, we apply this method to the analysis and stochastic characterization of general linear discrete systems with Gaussian input and measurement noise, and subjected to parameter discontinuity in any parameter or set of parameters of controlled process. The stochastic dynamic behavior following the discontinuity is studied in the plant state, filter and control strategies, and performance measure. This analysis may be specialized in characterizing the stochastic dynamic behavior of a stationary discrete system with arbitrary inputs and perfect measurements.

EXAMPLES

Although our theoretical and experimental investigations are under way, here we present some examples and their simulation results to illustrate the interesting nature of the observations. The results were obtained on the digital simulation of the process using 1000 simulations.

Example 1: First-Order Process

Consider a first-order process with sinusoidal input disturbance

$$x_p(k+1) = \alpha_p x_p(k) + \sin(0.5k + \phi) \tag{9}$$

where $\alpha_p = \alpha_0 + \Delta\alpha \, \mu_\alpha(k)$, and ϕ is a random phase. It is assumed that the parameter discontinuity occurs at k=0, and that the process was operating under steady-state conditions for k<0.

Figure 1 shows $E[x_p^2(k)]$ following the parameter discontinuity with the following two sets of parameters:

(a) $\alpha_0 = -0.5$, $\Delta\alpha = 1.0$,

(b) $\alpha_0 = 0.5$, $\Delta\alpha = -1$.

It is interesting to note the dynamic behavior in $E[x_p^2(k)]$ of the process for two different sets of parameters. Curve (a) shows a slight overshoot whereas curve (b) exhibits a number of oscillations which take about 6 seconds to settle down.

Example 2: Second-Order Process

In this example ~~the~~ a second-order process is considered, and is given by

$$\begin{bmatrix} x_{p1}(k+1) \\ x_{p2}(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -0.5 & \alpha_p \end{bmatrix} \begin{bmatrix} x_{p1}(k) \\ x_{p2}(k) \end{bmatrix} + \begin{bmatrix} 0 \\ w(k) \end{bmatrix} \tag{10}$$

where, as in Example 1,

$$\alpha_p = \alpha_0 + \Delta\alpha \, \mu_\alpha(k) \; .$$

Figure 2 shows the dynamic behavior of the process in $E[x_{p1}^2(k)]$ with a sinusoidal input disturbance and various values of nominal parameter α_0 and parameter discontinuity $\Delta\alpha$. Again the dynamic behavior of the process in $E[x_{p1}^2(k)]$ is interesting to note. Following the parameter discontinuity at k=0 the process exhibits large oscillations. Curve (d) shows the behavior of the process which can be modeled by an equivalent non-minimum phase type mathematical model.

Figure 3 illustrates the results of the above second-order process with zero mean random input disturbance. Curves (a) and (b) show that the steady-state values of $E[x_1^2(k)]$ and $E[x_{p1}^2(k)]$ are nearly the same, but following the parameter discontinuity there are large dynamic variations.

CONCLUSIONS

The problem of parameter discontinuity is studied in optimally controlled stochastic discrete process. Although only preliminary results are reported many important observations were made. For example, in a stationary situation, the steady-state stochastic behavior of the plant state before and after the parameter discontinuity may nearly be at the same level, but the parameter discontinuity may introduce violent dynamic transients - settling down to steady-state in several time-constants.

It seems that these observations are very important in the design of sub-systems which are most susceptable to parameter failures. This problem of parameter discontinuity, unfortunately, cannot be solved using the classical sensitivity methods, since a change in the parameter may not bring noticeable changes in the steady-state behavior but it may create large transients. It is envisaged that for small parameter discontinuity, one may obtain a linear mathematical model of the process. Further investigations in these directions are under way.

ACKNOWLEDGMENT

This work was supported by the National Research Council of Canada under Grants A-5625, and A-1080, and the Defence Research Board of Canada under Grants 4003-02 and 9781-04.

REFERENCES

1. D.D. Sworder, "Feedback Control of a Class of Linear Systems with Jump Parameters", IEEE Trans. Automatic Control, Vol. AC-14, pp. 9-14, Feb. 1969.
2. M.M. Gupta and P.N. Nikiforuk, "Some Further Contributions to the Dynamic Sensitivity of the Parameter Perturbation Processes", contributed to the Automatic Control Division for the Presentation at the 1970 ASME Winter Annual Meeting held at New York. Also to appear in the Trans. of the ASME Dynamic Systems, Measurements and Control, 1971.
3. A.V. Phatak, and G.A. Bekey, "Model of the Adaptive Behavior of the Human Operator in Response to a Sudden Change in the Control Situation", IEEE Trans. in Man-Machine Systems, Vol. MMS-10, pp. 72-80, Sept. 1969.
4. D.H. Weir and A.V. Phatak, "Model of Human Operator Response to Step Transitions in Controlled Element Dynamics", NASA CR-671, Jan. 1967.
5. D.H. Weir and W.A. Johnson, "Pilot Dynamic Response to Sudden Flight Control System Failures and Implication for Design", NASA CR-1087, June, 1968.
6. M.M. Gupta, P.N. Nikiforuk, and J.M. Milne, "Jump Type Behavioral Uncertainties in Stochastic Optimal Control Problems", Proceedings of the 1971 Joint Automatic Control Conference, St. Louis, August 11-13, 1971.

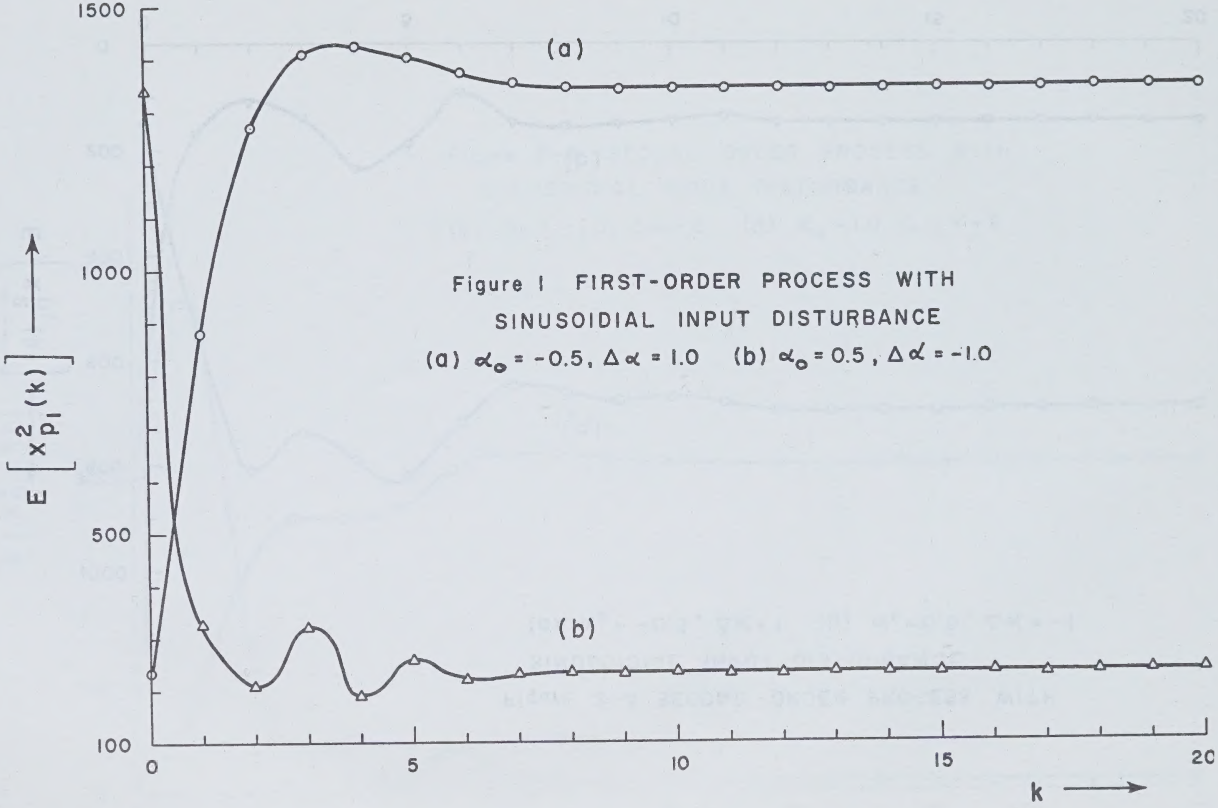


Figure 1 FIRST-ORDER PROCESS WITH SINUSOIDIAL INPUT DISTURBANCE

(a) $\alpha_o = -0.5$, $\Delta\alpha = 1.0$ (b) $\alpha_o = 0.5$, $\Delta\alpha = -1.0$

Figure 2-A SECOND-ORDER PROCESS WITH SINUSOIDIAL INPUT DISTURBANCE
(a) $\alpha_0 = -0.5$, $\Delta\alpha = 1$ (b) $\alpha_0 = 0.5$, $\Delta\alpha = -1$

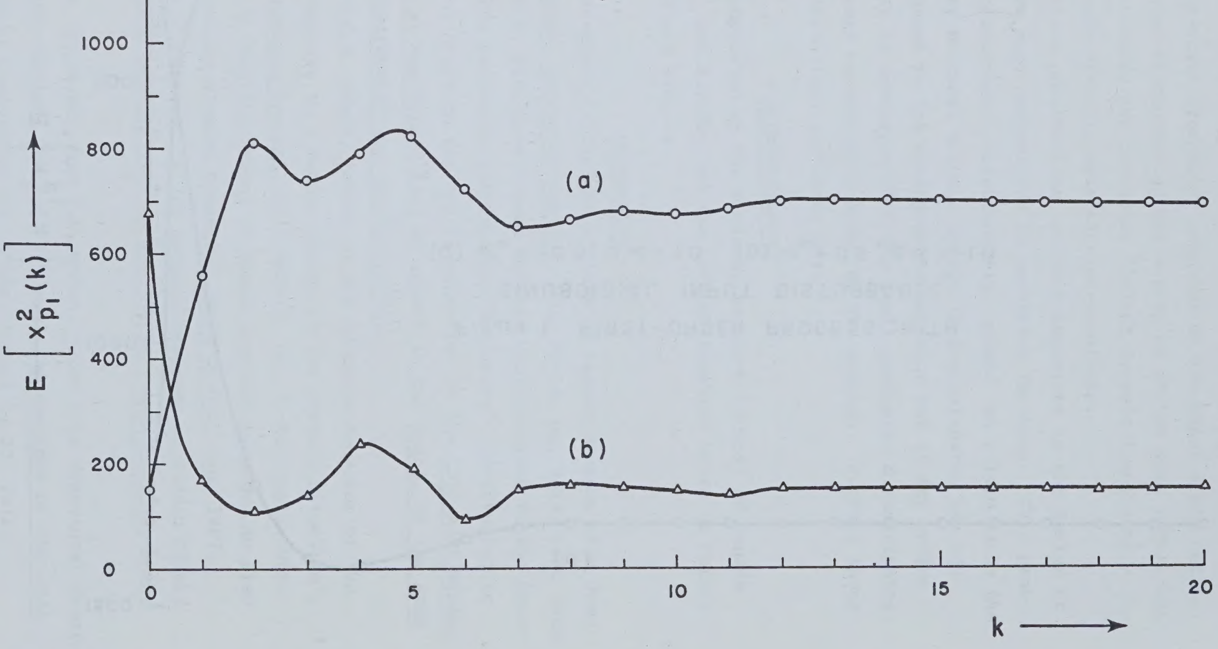
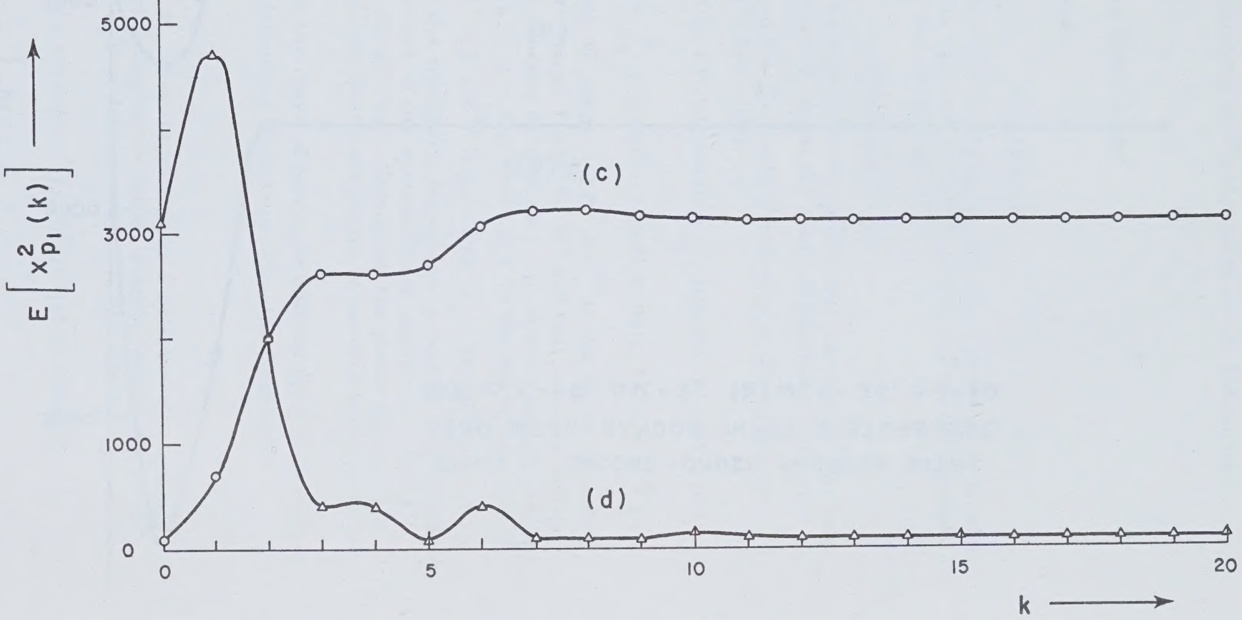


Figure 2-B SECOND-ORDER PROCESS WITH SINUSOIDIAL INPUT DISTURBANCE

(c) $\alpha_0 = -1.0$, $\Delta\alpha = 2$ (d) $\alpha_0 = 1.0$, $\Delta\alpha = -2$



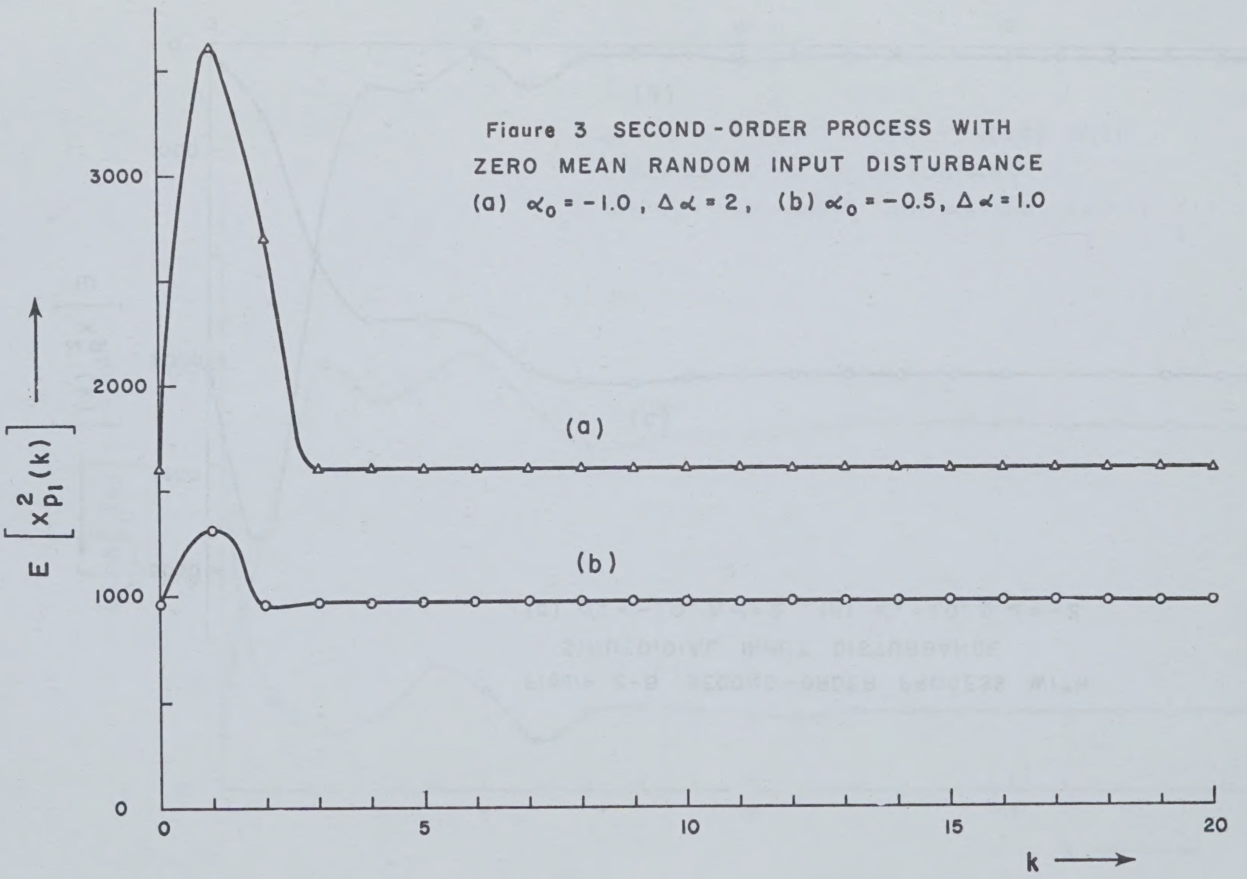


Figure 3 SECOND-ORDER PROCESS WITH ZERO MEAN RANDOM INPUT DISTURBANCE (a) $\alpha_0 = -1.0$, $\Delta\alpha = 2$, (b) $\alpha_0 = -0.5$, $\Delta\alpha = 1.0$

GENERALIZED STOCHASTIC OPTIMIZATION WITH APPLICATIONS TO STOCHASTIC CONTROL AND OPTIMAL DESIGN OF RANDOM SYSTEMS.

N. U. AHMED
Electrical Engineering Department
University of Ottawa,
Ottawa, Ontario. Canada.

ABSTRACT

In this paper a generalized stochastic optimization problem is solved. Several examples indicating application to stochastic control and optimal design of random systems are presented.

INTRODUCTION

Let I be any closed bounded interval of the real line and R^n the real Euclidean n-space. Consider $\hat{F}$ to be the class of functions $\{F(t,x)\}$ defined on $I \otimes R^n$ to R^n and let $\hat{S}$ denote the family of stochastic Ito-Differential systems in R^n described by

$$\hat{S}: \quad \{ dx(t) = F(t,x(t))dt + G(t,x(t),w(t))dw(t) + \int_{S^1} H(t,x(t),v)\, p(dt \otimes dv)$$

$$+ \int_{S_1} H(t,\, x(t),\, v)\, q(dt \otimes dv), \quad t \in I,\ F \in \hat{F} \} .$$

It is assumed thoughout that G maps $I \otimes R^n \otimes R^m$ to the space of (nxm) real valued matrices and H maps $I \otimes R^n$ to R^n and that they are fixed. w is an m-vector Wiener process defined on I, p is the Poisson random measure on $I \otimes R^n$ and $q \triangleq p - \mathcal{E}\, p$ where $\mathcal{E}\, p$ is the expected value of the random measure p. $S^1 = \{v \in R^n : |v| > 1$ and $S_1 = \{v \in R^n : |v| \leq 1\}$. For detailed properties of the stochastic measures p and q (contributing to the jumps of the trajectories of the system $\hat{S}$) the reader is referred to (Skorokhod [13], pp. 15, pp. 34). Thoughout the paper the notation $|\cdot|$ is used to denote the usual Euclidean length of vectors and matrices of finite dimension.

Let $D_1 \triangleq D_1(I,R^n)$ be the class of functions defined on I with values in R^n and admitting discontinuities of no more than the first kind. We need the following

Definition 1

A process $x^* = \{x^*(t),\ t \in I\}$ is said to be a solution of a system from the class $\hat{S}$, corresponding to the initial random vector x^o, if and only if there exists an $F^* \in \hat{F}$ so that

$$x^*(t) = x^o + \int_{t_o}^{t} F^*(\tau, x^*(\tau))d\tau + \int_{t_o}^{t} G(\tau, x^*(\tau), w(\tau))dw(\tau)$$

$$+ \int_{t_o}^{t} \int_{S^1} H(\tau, x^*(\tau), v)p(d\tau \otimes dv) + \int_{t_o}^{t} \int_{S_1} H(\tau, x^*(\tau), v)q(d\tau \otimes dv)$$

for each $t \in I$ with probability one (P - a.e) and $x^* \in D_1$ P - a.e. .

The optimization problem considered in this paper may now be stated as follows: Consider $\varphi: \hat{F} \otimes D_1 \to R$ to be a real valued non-negative lower-semicontinuous functional. The problem is to choose a system $S^* \in \hat{S}$ in other words an $F^* \in \hat{F}$ so that the cost functional $J(F^*) \triangleq \mathcal{E}\,\varphi(F^*, x^*)$ evaluated along the trajectory x^* corresponding to the system S^* is a minimum. That is, $J(F^*) \leq J(F)$ for all $F \in \hat{F}$. This problem is solved in propositions 4 and 5 of section 3 after necessary materials are developed in section 2.

The following notations and terminology will be used throughout the paper without repeating their significance: B_o denotes the smallest σ-algebra with respect to which the random initial vector x^o is measurable. For each pair $t, \tau \in I \triangleq [t_o, t_1]$ and $t < \tau$, $B_{t,\tau}(z)$ denotes the smallest σ-algebra with respect to which the events $\{z(s), t \leq s \leq \tau\}$ corresponding to the stochastic (vector) process $\{z(t), t \in I\}$ are measurable. Similarly $B_{t,\tau}(p)$ denotes the least σ-algebra with respect to which the events $\{p([s,r] \otimes A); t_o \leq s \leq r \leq t_1, A \subset R^n\}$ corresponding to the random Poisson measure p are measurable; where A is any Borel set of R^n. We are mainly concerned with the σ-algebras $B_{t,\tau}(w)$, $B_{t,\tau}(dw)$, $B_{t,\tau}(p)$ and $B_{t,\tau}(x)$ that naturally arise in the study of the class of stochastic systems $\hat{S}$.

Let us consider the probability (measure) space (Ω, B, P) where Ω is the supporting set of the triple (x^o, w, p) and $B \triangleq B_{t_o;t_1} \triangleq B_o \vee B_{t_o,t_1}(w) \vee B_{t_o,t_1}(p)$ is the smallest σ-algebra with respect to which the triple (x^o, w, p) is measurable and P is the probability measure with $P(\Omega) = 1$. Further, for every pair $\tau, s \in I$, $B_{\tau,s} \subset B$ represents the σ-algebra of events defined on the interval $[\tau, s] \subset I$.

Without any loss of generality we may assume that $P\{w(t_o) = 0\} = 1$ and that $\mathcal{E}\,|w_i(t)|^2 = \lambda_i\,|t - t_o|$, $\lambda_i > 0$ for $i = 1, 2, \ldots m$ where w_i is the ith component of the m-vector Wiener process w defined on I.

BASIC ASSUMPTIONS

It is assumed throughout the paper that $\hat{F}$ possesses the following properties:

$A_{\hat{F}}$(i): there exists a $K \in L_2(I)$ so that for all $F \in \hat{F}$

$|F(t,x) - F(t,y)| \leq |K(t)|\,|x-y|$ a.e. on I for all (x,y) in any bounded region in R^n.

$A_{\hat{F}}$(ii): there exists $\alpha, \beta > 0$ so that for all $F \in \hat{F}$ $|F(t,x)| \leq (\alpha + \beta\,|x|)\,|K(t)|$ a.e. on I for all x in any bounded region in R^n.

$A_{\hat{F}}$(iii): $\lim\limits_{|h| \to o} \int_I |F(t+h,x) - F(t,x)|^2\,dt = 0$ uniformly with respect to $F \in \hat{F}$ for every fixed $x \in R^n$.

For G and H we introduce the following assumptions:

$A_G(i)$: there exists a function L_G on $I \otimes R^m$ to R so that for every $z \in R^m$ it is Lebesgue measurable in t on I and for almost all $t \in I$ it is continuous in z on R^m and further $\int_I \mathcal{E} \, |L_G(t, w(t))|^2 \, dt < \infty$ for the given Wiener process in R^m and that

$$|G(t,x,w(t)) - G(t,y, w(t))| \leq |L_G(t, w(t))| \; |x-y|$$

$\mu \cdot P$ a.e on $I \otimes \Omega$ for all $x, y \in R^n$. Here μ stands for the Lebesgue measure on I and P for the probability measure and $\mathcal{E} \, h \triangleq \int_\Omega h dP$ denotes the mathematical expectation of the random variable h.

$A_G(ii)$: there exists $\alpha_G, \beta_G, \gamma_G > 0$ and an integer $s \geq 1$ and a function $K_G \in L_2(I)$ so that

$$|G(t,x,w(t))| \leq |K_G(t)| \; [\alpha_G + \beta_G |x|] + \gamma_G |w(t)|^s$$

for all x in a bounded region in R^n.

$A_H(i)$: there exists a function $K_H \in L_2(I)$ so that for all x, y in a bounded region in R_n

$$\left(\int_{R^n} |H(t,x,v) - H(t,y,v)|^2 \frac{dv}{|v|^{n+1}}\right)^{1/2} \leq |K_H(t)| \; |x-y| \quad \text{a.e on } I$$

and

$A_H(ii)$: there exists $\alpha_H, \beta_H > 0$ so that

$$\left(\int_{R^n} |H(t,x,v)|^2 \frac{dv}{|v|^{n+1}}\right)^{1/2} \leq |K_H(t)| \; [\alpha_H + \beta_H |x|]$$

for all x in a bounded region in R^n.

A_o : $\mathcal{E} \, |x^o|^2 \leq c_o^2 < \infty$.

2. EXISTENCE AND UNIQUENESS OF SOLUTION.

Lemma 1.

Suppose the processes $x \triangleq \{x(t), t \in I\}$ and $y \triangleq \{y(t), t \in I\}$ are B measurable and belong to D_1 with probability one and further for a fixed but almost all $t \in I$ the random vectors $x(t)$ and $y(t)$ are $B_{t_o, t}$ measurable and independent of the events corresponding to the σ-algebra $B_{\tau, s}$ for $t < \tau \leq s \leq t_1$. Then if the matrix valued function G satisfies the property $A_G(i)$

$$\mathcal{E} \, |G(t,x(t), w(t)) - G(t,y(t), w(t))|^2 \leq \mathcal{E} \, |L_G(t, w(t))|^2 \; \mathcal{E} \, |x(t)-y(t)|^2$$

for almost all $t \in I$. 2.1

Proof: For detailed proof see (Ahmed 3, 1971).

Proposition 1.

Under the assumptions $A_{\hat{F}}$, A_G, A_H and A_o and for every fixed

$F \in \hat{F}$ the corresponding system $S_F \in \hat{S}$ has a unique solution x which belongs to D_1 with probability one and which has a bounded second moment on every finite interval $I \subset [t_o, \infty)$ and furthermore $B_{t_o,t}(x) \subset B_o \vee B_{t_o,t}(w) \vee B_{t_o,t}(p)$.

Proof: For detailed proof see (Ahmed 3, 1971).

Definition 2

Let us introduce a class M of stochastic processes defined on $I \subset R$ with values in R^n and possessing the following properties:

(i) $x \in D_1$ P-a.e.

(ii) at every point $t \in I$, except the terminal point, $x(t)$ is defined by its value at $t + o$, that is, $x(t) \triangleq x(t+o)$.

(iii) $\sup_{t \in I} \mathcal{E} |x(t)|^2 < \infty$ where $|z| \triangleq \sqrt{\Sigma |z_i|^2}$.

With the above definition for M it is easily verified that M is a Banach space with the norm $\|x\| \triangleq \sup_{t \in I} \sqrt{\mathcal{E} |x(t)|^2}$ and consequently it is also a complete metric space with the metric $\rho(x,y) \triangleq \sup_{t \in I} \sqrt{\mathcal{E} |x(t) - y(t)|^2}$

Remark 1 Thus it follows from the proposition 1 that the solution x of the system $S_F \in \hat{S}$ corresponding to the element $F \in \hat{F}$ and the initial random vector x^o (satisfying the basic assumption A_o) belongs to the B-space M. This solution process may be denoted by x_F.

Definition 3

The family of stochastic processes $X \triangleq \{x_F | F \in \hat{F}\}$ generated by the trajectories of the class of systems $\hat{S}$ is called the "solution manifold ".

Remark 2 It follows from proposition 1 and the definition 2 and 3 that $X \subset M$.

The following lemma will be found useful in the sequel.

Lemma 2

The set X is a bounded subset of the B-space M.

Remark 3

From the above lemma it follows that for every $\epsilon > 0$ there exists a compact set $E_\epsilon \subset R^n$ so that $P\{x_F(t) \in E_\epsilon$ for all $t \in I$ and all $F \in \hat{F}\} > 1-\epsilon$. Let us define the class

$$\hat{F}_\epsilon = \{F \in \hat{F} : F(t,x) = 0 \ \ \mu\text{-a.e for } x \notin E_\epsilon\}.$$

Thus, for all $F \in \hat{F}_\epsilon$, it follows from $A_{\hat{F}}$ (ii) that, $|F(t,x)| \le a |K(t)|$ a.e. where $a = \alpha + \beta \sup_{x \in E_\epsilon} |x|$ and consequently $\hat{F}_\epsilon$ is a bounded subset of the $L_2(I, R^n)$ space. That is, $\int_I |F(t,x)|^2 dt \le |a|^2 \|K\|^2$ for all $F \in \hat{F}_\epsilon$ and for all $x \in R^n$.

The following lemma is taken from (Dunford [4], pp. 298, Th. 20) for reference.

Lemma 3

Let I be any finite interval on the real line and consider the $L_2(I, R^n)$ space. A subset $B \subset L_2(I, R^n)$ is conditionally compact (i.e $\overline{B}$ is compact) iff

(i) B is bounded.

and (ii) $\lim_{|h| \to 0} \int_I |f(t+h) - f(t)|^2 \, dt = 0$ uniformly with respect to $f \in B$.

Remark 4

From the above lemma and the conditions $A_{\hat{F}}$(ii), $A_{\hat{F}}$(iii) it follows that the set $\hat{F}_\epsilon$ is a conditionally compact subset of $L_2(I, R^n)$. Define $\tilde{F}_\epsilon$ to be the strong closure of the set $\hat{F}_\epsilon$. Thus $\tilde{F}_\epsilon$, as a subset of the $L_2(I, R^n)$ space, is (strongly) compact (and closed). Define

$X_\epsilon = \{x_F : F \in \tilde{F}_\epsilon\}$.

Proposition 2

The "solution manifold" X is a sequentially compact subset of the B-space M.

Proof: For detailed proof see (Ahmed 3, 1971).

Corollary 1.

The sets X and X_ϵ $(\epsilon > 0)$ are conditionally compact subsets of the complete metric space M.

Proof;

The proof follows from proposition 2 and the wellknown result (Dunford [4], pp. 22, Th. 15) that in a metric space a sequentially compact subset is conditionally compact. Thus both $\overline{X}$ and $\overline{X}_\epsilon$ for $\epsilon > 0$ are compact.

3. QUASI OPTIMAL SOLUTIONS

It may be noted that the solution manifold X, though a sequentially compact subset of the B-space M, is not necessarily closed. Therefore the optimization problem stated at the beginning of the first section may not have a solution. In this section we intend to prove the existence of a quasi optimal solution. For this purpose we need the following

Proposition 3.

The set $X_\epsilon = \{x_F : F \in \tilde{F}_\epsilon\} \subset X$ is a compact subset of the B-space M.

Proof:

For detailed proof see (Ahmed 3, 1971).

The following result will be found useful in the proof of existence of a quasi optimal (ϵ-optimal) solution (definition 4) to the optimization problem stated in the introduction.

Proposition 4

Suppose the triple (F,G,H) satisfy the basic assumptions $A_{\hat{F}}$, A_G and A_H. Consider the family of Ito-differential systems $\hat{S}$ with $F \in \tilde{F}_\epsilon$ where $\tilde{F}_\epsilon$ is the strong closure (in $L^2(I,R^n)$) of $\hat{F}_\epsilon \triangleq \{F \in \hat{F}: F(t,x) = 0\mu$ - a.e for all $x \notin E_\epsilon$, E_ϵ - a compact subset of $R^n\}$. Let $\varphi : L_2(I,R^n) \otimes D_1(I,R^n) \rightarrow R$ be a real valued non-negative functional lower semicontinuous in the strong topology of $L_2(I,R^n)$ and in the uniform topology of $D_1(I,R^n)$.

Then there exists an $F_o \in \tilde{F}_\epsilon$ and $x_o \in X_\epsilon$ with $x_o = S_{F_o} x_o$ P-a.e so that $\mathcal{E}\varphi(F_o,x_o) \le \mathcal{E}\varphi(F,x)$ for all pairs $(F,x) \in Y \triangleq \{F,x) \mid x = S_F x$ P-a.e, $F \in \tilde{F}_\epsilon\}$.

Definition 4.

If for an arbitrary but fixed $\gamma > 0$ there exists an $F^* \in \hat{F}$ so that $J(F^*) \le \inf_{F \in \hat{F}} J(F) + \gamma$ then the corresponding system $S_{F^*} \in \hat{S}$ is said to be γ-optimal.

Using this definition and the previous results we obtain the following result.

Proposition 5.

Consider the family of stochastic dynamical systems $\hat{S}$ and suppose (i) the basic assumptions $A_{\hat{F}}$, A_G, A_H and A_o hold; (ii) $f^o(t,\xi,\eta)$, defined on $I \otimes R^n \otimes R^n$, is a non-negative measurable function of t on I for each pair $(\xi,\eta) \in R^n \otimes R^n$ and jointly continuous in $(\xi,\eta) \in R^n \otimes R^n$ for almost all $t \in I$; and (iii) there exist a Lebesgue measurable function $K_o \in L_1(I)$ and numbers $0 \le a_1, a_2 < \infty$ so that $|f^o(t,\xi,\eta)| \le |K_o(t)| + a_1|\xi|^2 + a_2|\eta|^2$ a.e on I for all $(\xi,\eta) \in R^n \otimes R^n$.

Then for every $\gamma > 0$ there exists an $F^* \in \hat{F}$ so that the cost functional $J(F^*) \triangleq \mathcal{E}\int_I f^o\,(t\ F^*(t,x^*(t)),\ x^*(t))\,dt$ evaluated along the trajectory x^* of the system S_{F^*} satisfies the inequality $J(F^*) \le \inf_{F \in \hat{F}} J(F) + \gamma$.

Proof:

For detailed proof see (Ahmed 3, 1971).

Remarks:

(a) The above result signifies that even when there is no optimal system in the class $\hat{S}$ there is at least an approximately optimal one i.e there is a quasi-optimal policy.

(b) It is interesting to mention that if the function f^o appearing in the definition of the functional $J(F)$ satisfies a Lipschitz condition $|f^o(t, y_o, x_o) - f^o(t, y_1, x_1)| \leq a|y_o - y_1| + b|x_o - x_1|$ $t \in I$, $0 < a, b < \infty$, then the problem can be reduced to a terminal control problem.

$$S': \begin{cases} dx(t) = F\,dt + G\,dw + \int_{S^1} H\, p(dt \otimes dv) + \int_{S_1} H\, q(dt \otimes dv) \\ dx^o(t) = f^o(t, F(t, x(t)), x(t))dt. \qquad t \in I \end{cases}$$

It is required to find an $F \in \hat{F}$ so that $J(F) = \mathcal{E}(x^o(t_1))$ is a minimum. Since X'_ϵ corresponding to the augmented system S' is compact the section set $\{x'(t_1) \mid x' \in X'_\epsilon\}$ is also compact. Thus $J(F) = \mathcal{E}\varphi(x'(t_1)) \triangleq \mathcal{E}(x^o(t_1))$ has a minimum in $\widetilde{F}_\epsilon$. Hence, by proposition 5, there exists a quasi-optimal system.

4. APPLICATIONS TO STOCHASTIC CONTROL

In this section we present few examples of application to optimal control of stochastic systems.

Let U be a compact subset of R^m and $\sigma(U)$ the σ-algebra of Borel sets in U. Let Θ denote the complete metric space of bounded Borel measures with domain $\sigma(U)$. Let I be any closed bounded subset of the real line and let N denote the space of measure valued functions on I with range in Θ i.e for each $\theta \in N$ and $t \in I$ $\theta_t \in \Theta$ and $\theta_t(A) < \infty$ for each $A \in \sigma(U)$.

Definition 5.

A subset $N_o \subset N$ is said to be weakly sequentially compact if for every sequence $\{\theta^n\} \in N_o$ there exists a subsequence $\{\theta^{n_k}\}$ and an element $\theta \in N_o$ so that for each $g \in C(U, R^n)$ the integral

$$\int_U g(u)\, \theta_t^{n_k}(du) \rightarrow \int_U g(u)\, \theta_t(du) \text{ for each } t \in I.$$

Definition 6.

A subset $N_o \subset N$ is said to be weakly equicontinuous if for each t in the interior of the set I and for each $g \in C(U, R^n)$

$$\lim_{h \to o} \int g(u)\, \theta_{t+h}(du) \rightarrow \int g(u)\, \theta_t(du) \text{ uniformly with respect to } \theta \in N_o.$$

Example 1

Let N_o be a weakly sequentially compact and weakly equicontinuous subset of the space of measures N. Consider the system T_2 described by the following stochastic Ito-differential equation

$$T_2: \{ dx(t) = \left(\int_U f(t, x(t), u)\, \theta_t(du) \right)dt + G\,dw + \int_{S^1} H p(dt \otimes dv) + \int_{S_1} H\, q(dt \otimes dv)$$

$t \in I$, $\theta \in N_o$,

where all the terms in the right hand side except the first remain unchanged.

The function $f: I \otimes R^n \otimes U \to R^n$ is assumed to satisfy the properties (i), (ii) and (iii) of example 1 (Ahmed 3, 1971). For each $(t,x) \in I \otimes R^n$ define the set valued function $R(t,x)$ by

$$R(t,x) \triangleq \{\int_U f(t,x,u)\, \theta_t(du),\ \theta \in N_o\} \text{ and define}$$

$$\hat{F}_2 \triangleq \{F: I \otimes R^n \to R^n \mid F(t,x) \in R(t,x)\} .$$

By use of the result (a) of lemma 4,(Ahmed 3), Lebesgue dominated convergence theorem, the assumptions (i), (ii) and (iii), and the assumed properties for the family of measures N_o it can be shown that the set $\hat{F}_2$, as defined above, satisfies the basic assumption $A_{\hat{F}}$. Consequently the system T_2 is also a special case. Thus the results of the preceeding sections are valid for this system. Further, as in example 1,(Ahmed 3)there is an optimal system and not merely a quasi-optimal one.

Example 2.

Let $F_o(t,x)$ be a given function on $I \otimes R^n$ to R^n satisfying the properties $A_{\hat{F}}$(i) and $A_{\hat{F}}$(ii). Denote by Q the metric space of $(n \times s)$ matrices $(s \le n)$ with the metric $|q^1 - q^2| = (\sum_{ij} |q^1_{ij} - q^2_{ij}|^2)^{1/2}$ and by $L_2(I,Q)$ the space of Lebesgue square integrable functions on I with values in the metric space Q.

Let $\hat{A}$ be a closed subset of $L_2(I,Q)$ with the properties

(i) $\|A\|^2 = \int_I |A(t)|^2\, dt \le c < \infty$ for all $A \in \hat{A}$ and

(ii) $\lim_{h \to o} \int_I |A(t+h) - A(t)|^2\, dt = 0$ uniformly with respect to $A \in \hat{A}$.

These conditions imply that $\hat{A}$ is a compact subset $L_2(I,Q)$.

Consider the stochastic differential equation

$$T_4 \begin{cases} dx(t) = F_o(t,x(t))\, dt + u(t)\, dt + G dw + \int_{S^1} Hp(dt \otimes dv) + \int_{S_1} Hq(dt \otimes dv) \\ u(t) = A(t)\, x'(t), \quad t \in I, \quad A \in \hat{A} , \end{cases}$$

where x' is any of the s-components of the vector x as before. The problem is to choose a feedback gain matrix $A \in \hat{A}$ that minimizes the cost functional $J(A) = \mathcal{E} \int_I f^o(t, A(t), x(t))\, dt$ where x is the response of the system T_4 corresponding to the matrix $A \in \hat{A}$ and the given initial random vector x^o i.e $x(t) = x(t,A,x^o)$, $t \in I$.

It is easily shown that a sufficient condition for the existence of an optimal feedback gain $A^* \in \hat{A}$ is that the functional $\varphi(A,x) = \int_I f^o(t, A(t), x(t))\, dt$ is lower semicontinuous on $L_2(I,Q) \otimes D_1(I,R^n)$ where $f^o : I \otimes Q \otimes R^n \to R$.

CONCLUSION

In this paper we considered the problem of optimal design of a large class of stochastic differential systems whose solutions are Markov processes with jumps. It was assumed that optimization could be carried out by a suitable choice of the infinitesimal drift coefficient F alone. It is, however, interesting to mention that the results of this paper can be easily extended to include the optimal choice of all the three variables F, G, and H instead of F alone. The results of this paper can be extended to include hereditary stochastic dynamical systems also, (Fleming and Nisio [6], pp. 777-794; Ahmed [1], pp. 45-53).

It appears to the author that it would be interesting to consider $\hat{F}$ as a subset of $L_1(I, R^n)$ instead of merely $L_2(I, R^n)$.

REFERENCES

1. Ahmed, N.U., "Optimal Control of a large class of stochastic differential systems with finite memory", Proc. 7th Ann. Allerton Conf., pp. 44-53, (1969).
2. Ahmed, N.U., "Closure and completeness of Wiener's orthogonal set $\{G_n\}$ in the class $L^2(\Omega, B, \mu)$ and its application to stochastic hereditary differential systems", J. Inf. and Control, Vol 17, No. 2, pp. 161-174, September (1970).
3. Ahmed, N.U., "Solution of Generalized Stochastic Optimization Problems" T.R. 71-3, Dept. of E.E., Univ. of Ottawa, 1971.
4. Dunford, N., and Schwartz, J.T., "Linear operator, part 1", Interscience, pp. 22, pp. 298, (1964).
5. Fleming, W.H., "Optimal control of partially observable diffusions", SIAM J. Control, Vol. 6, No. 2, pp. 194-214, (1968).
6. Fleming, W.H., and Nisio, M., "On the existence of optimal stochastic controls", J. Math. Mech., Vol. 15, pp. 777-794, (1966).
7. Skorokhod, A.V., "Studies in the theory of random processes" Addison-Wesley Publishing Company, Inc., pp. 15, pp. 16, pp. 34, (1965).

This work was supported by the National Research Council of Canada, under grant No. A-7109.

OPTIMAL CONTROL OF NON-UNIFORMLY SAMPLED DISCRETE SYSTEMS BY NONLINEAR PROGRAMMING *

L. R. NARDIZZI AND B. K. BHAGAVAN
Information & Control Sciences Center, Institute of Tech.
Southern Methodist University, Dallas, Texas 75222

ABSTRACT

Linear and nonlinear optimal control problems with state and control inequality constraints are converted into equivalent parameter optimization problems with piecewise constant control which are then solved subject to the original constraints using a nonlinear programming technique. Non-uniformly sampled systems with explicit constraints on sampling periods in addition to other state and control inequality constraints are conveniently handled. A few examples illustrate that the method is equally effective for linear and nonlinear dynamics.

I. INTRODUCTION

In the optimal control of discrete-time systems where the inputs are piecewise constant, a non-uniform sampling (switching) scheme is desirable. Thus the problem is to optimize the cost functional with respect to the switching times and the input amplitudes. Though the discrete maximum principle can be used to solve such a problem, the computational difficulties involved prohibit its use, especially where constraints on sampling periods are involved (1). Though, theoretically, dynamic programming can be employed, its use is mainly limited by the amount of computer storage required.

At present the best way and perhaps the only reasonable way to obtain computational results for the optimization of a discrete problem with constraints on states, input amplitudes and the sampling periods is by the use of nonlinear programming techniques.

II. FORMULATION OF THE PROBLEM

(1) The dynamics of the plant are governed by

$$\underline{x}(i+1) = f[\underline{x}(i), \underline{u}(i), i] \tag{1}$$

where $\underline{f}$ and $\underline{x}$ are n-dimensional vectors; $\underline{u}$ is an m-dimensional vector; $i = 1,2, \ldots, (N-1)$; N = number of switchings (sampling periods). The control $\underline{u}(i)$ has the form

$$\underline{u}(i) = \eta(T_i) \begin{bmatrix} u_1(i) \\ \vdots \\ u_m(i) \end{bmatrix} \qquad \text{where } \eta(T_i) = 1 \text{ during the i-th sampling duration} \\ = 0 \text{ otherwise}$$

---

*This work was supported in part by the National Science Foundation under Grant No. NSF GK-5608.

If the given problem has continuous dynamics, a convenient discretization scheme is employed to convert the continuous system into (1).

(2) N, the number of switchings is fixed.

(3) $\underline{x}(0) = \underline{x}_0$ is specified.

(4) $\underline{x}(N) = \underline{x}_f$ may be specified.

(5) It is required to minimize a cost function

$$J = \sum_{i=1}^{N-1} \Phi[\underline{x}(i), \underline{u}(i), i] \qquad (2)$$

If the cost is of an integral type for continuous dynamics, it is reduced to the form of eq. (2) by the same scheme used to discretize the system equations.

(6) The duration between samplings, T_i, is to be optimized between specified limits

$$T_{min} \le T_i \le T_{max} \qquad i = 1,2, \ldots\ldots, N \qquad (3)$$

(7) Upper and lower bounds on the amplitudes of each component of the control vector are specified.

$$u_j(i)_{min} \le u_j(i) \le u_j(i)_{max} \qquad \begin{array}{l} i = 1,2, \ldots., N \\ j = 1,2, \ldots., m \end{array} \qquad (4)$$

(8) Inequality constraints on state and control variables of the type,

$$C(k)\ (\underline{x}(i),\underline{u}(k)) \ge 0 \qquad k = 1,2, \ldots., K \qquad (5)$$

can also be specified.

The problem is to find a set of control amplitudes $\underline{u}(j)$, and a set of sampling intervals T_j, $j = 1,2, \ldots., N$, to minimize a cost given by eq. (2) subject to the constraints given by equations 3, 4 and 5.

III. THE COMPUTATIONAL TECHNIQUE

The modified cost for the system given by (1) and cost (2) would be

$$J' = \sum_{i=1}^{N-1} \Phi[\underline{x}(i),\underline{u}(i),i] + K||\underline{x}(N)-\underline{x}_f||^2 \qquad (6)$$

where the value of $K > 0$, is sequentially increased until $\underline{x}(N)$ is sufficiently close to $\underline{x}_f$. Since piecewise constant controls are assumed the optimization parameters are the N-1 control amplitudes and the N-1 sampling intervals. Once the control amplitudes and sampling intervals are fixed, the states are readily calculated using the system dynamics. On the other hand, the states at every stage could be considered as independent variables and the system equation adjoined to the cost. This obviously increases the computer storage required and the computation time.

The computational technique used is a nonlinear programming approach with a multiple gradient summation algorithm suggested by Klingman and

Himmelblau (2). This is a modified direct search technique which can handle inequality constraints in a convenient fashion.

A general outline of the computational procedure is briefly discussed in the Appendix.

Since the independent variables of optimization are u(i) and T_i, i = 1,2, ..., N-1, it is necessary to find the partial derivatives of the cost and the inequality constraints with respect to these variables. Hence, it is required to evaluate

$$\frac{\partial J}{\partial T_i}, \quad \frac{\partial J}{\partial \underline{u}(i)}, \quad \frac{\partial C(k)}{\partial \underline{u}(i)} \quad \text{and} \quad \frac{\partial C(k)}{\partial T_i} \qquad \begin{array}{l} k = 1,2, \ldots, K \\ i = 1,2, \ldots, N-1. \end{array}$$

if the discrete dynamics are derived out of a continuous system which is linear in the state and time invariant such as

$$\dot{\underline{x}} = A\underline{x} + f(u) \tag{7}$$

then the discretized dynamics would be

$$\underline{x}(i+1) = \phi(T_{i+1})\underline{x}(i) + \int_0^{T_{i+1}} \phi(T_{i+1}-\sigma) f(\underline{u}(i+1))d\sigma \tag{8}$$

where $\phi(T_{i+1}) = e^{AT_{i+1}}$. Hence knowing Φ, analytical expressions for

$$\left[\frac{\partial \underline{x}(i+1)}{\partial T_{i+1}}\right], \quad \left[\frac{\partial \underline{x}(i+1)}{\partial \underline{u}(i+1)}\right], \quad \left[\frac{\partial \underline{x}(i+1)}{\partial \underline{x}(i)}\right]$$

could be developed where

$$\left[\frac{\partial \underline{x}(i+1)}{\partial T_{i+1}}\right] \text{ is a n-vector with } \left(\frac{\partial \underline{x}(i+1)}{\partial T_{i+1}}\right)_j = \frac{\partial x_j(i+1)}{\partial T_{i+1}}$$

$$\left[\frac{\partial \underline{x}(i+1)}{\partial \underline{u}(i+1)}\right] \text{ is an nxm matrix with } \left(\frac{\partial \underline{x}(i+1)}{\partial \underline{u}(i+1)}\right)_{j,k} = \frac{\partial x_j(i+1)}{\partial u_k(i+1)}$$

$$\left[\frac{\partial \underline{x}(i+1)}{\partial \underline{x}(i)}\right] \text{ is an nxn matrix with } \left(\frac{\partial \underline{x}(i+1)}{\partial \underline{x}(i)}\right)_{j,k} = \frac{\partial x_j(i+1)}{\partial x_k(i)}$$

The required partials can now be evaluated as follows:

$$\begin{aligned} \left[\frac{\partial \underline{x}(i)}{\partial T_j}\right] &= 0 && \text{for } j > i \\ &= \left[\frac{\partial \underline{x}(i)}{\partial T_i}\right] && \text{for } j = i \end{aligned} \tag{9}$$

$$= \left[\frac{\partial \underline{x}(i)}{\partial \underline{x}(i-1)}\right] \left[\frac{\partial \underline{x}(i-1)}{\partial T_j}\right] \qquad \text{for } j < i$$

Similarly,

$$\left[\frac{\partial \underline{x}(i)}{\partial \underline{u}(j)}\right] = 0 \qquad \text{for } j > i$$

$$= \left[\frac{\partial \underline{x}(i)}{\partial \underline{u}(i)}\right] \qquad \text{for } j = i \qquad (10)$$

$$= \left[\frac{\partial \underline{x}(i)}{\partial \underline{x}(i-1)}\right] \left[\frac{\partial \underline{x}(i-1)}{\partial \underline{u}(j)}\right] \qquad \text{for } j < i$$

Using these,

$$\frac{\partial J}{\partial T_i} = \sum_{j=1}^{N-1} \left[\frac{\partial J}{\partial \underline{x}(j)}\right] \left[\frac{\partial \underline{x}(j)}{\partial T_i}\right] + \left[\frac{\partial J}{\partial T_i}\right]^* \qquad (11)$$

$$\frac{\partial C(k)}{T_i} = \sum_{j=1}^{N-1} \left[\frac{\partial C(k)}{\partial \underline{x}(j)}\right] \left[\frac{\partial \underline{x}(j)}{\partial T_i}\right] \qquad (12)$$

$$\left[\frac{\partial J}{\partial \underline{u}(i)}\right] = \sum_{j=1}^{N-1} \left[\frac{\partial J}{\partial \underline{x}(j)}\right] \left[\frac{\partial \underline{x}(j)}{\partial \underline{u}(i)}\right] + \left[\frac{\partial J}{\partial u(i)}\right]^* \qquad (13)$$

$$\left[\frac{\partial C(k)}{\underline{u}(i)}\right] = \sum_{i=1}^{N-1} \left[\frac{\partial C(k)}{\partial \underline{x}(j)}\right] \left[\frac{\partial \underline{x}(i)}{\partial \underline{u}(i)}\right] \quad \left[\frac{\partial C(k)}{\partial u(i)}\right]^* \qquad (14)$$

where

$\left[\frac{\partial J}{\partial \underline{u}(i)}\right]$ is a 1xm matrix with $\left(\frac{\partial J}{\partial \underline{u}(i)}\right)_k = \frac{\partial J}{\partial u_k(i)}$

$\left[\frac{\partial J}{\partial \underline{x}(j)}\right]$ is a 1xn matrix with $\left(\frac{\partial J}{\partial \underline{x}(j)}\right)_k = \frac{\partial J}{\partial x_j(i)}$

$\left[\frac{\partial C(k)}{\partial \underline{u}(i)}\right]$ is a 1xm matrix with $\left(\frac{\partial C(k)}{\partial \underline{u}(i)}\right)_j = \frac{\partial C(k)}{\partial x_j(i)}$

$\left[\frac{\partial C(k)}{\partial \underline{x}(j)}\right]$ is a 1xn matrix with $\left(\frac{\partial C(k)}{\partial \underline{x}(i)}\right)_j = \frac{\partial C(k)}{\partial x_j(i)}$

and $\left[\frac{\partial f}{\partial y}\right]^*$ represents partial derivative of f with respect to y considering y as an independent variable. The above procedure can be used for linear continuous systems and discrete systems. However, for the more complicated case of nonlinear continuous systems, in the absence of an accurate discretization scheme, the partials are numerically computed.

Thus, knowing a set of values for $u_j(i)$ and T_i, one can readily compute the cost, the value of the constraints and all the required partial derivatives. The states at the switching instans and the cost of the original continuous system are computed by numerical integration using a digital computer. The method suggested in this paper is also amenable to hybrid computer solution if the dynamics and cost are continuous and the control is discrete. On the other hand if both the original system and control are in discrete form to begin with, then a digital computer solution is very convenient.

IV. EXAMPLES

Example 1: Four different cases are considered for the same plant dynamics. The system is the same as the one used by Kuo and Tabak (3).

The transfer function of the continuous linear system is

$$\frac{C(s)}{U(s)} = \frac{0.16}{s^2 + 0.032s + 0.16} \tag{15}$$

The state equations are

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -0.16x_1 - 0.032x_2 + 0.16u \end{aligned} \tag{16}$$

with $x_1(0) = x_2(0) = 0$ and $x_1(t_f) = 2$, $x_2(t_f) = 0$ and t_f is unspecified.

The discretized equations, assuming the control is constant between switchings, are

$$\begin{aligned} x_1(i+1) &= \exp(-0.016T_i)\,[\sin(0.4T_i + 1.531)x_1(i) + 2.5\sin(0.4T_i) \\ &\quad x_2(i)] + [1-\exp(-0.016T_i)\sin(0.4T_i + 1.531)]\,u(i) \\ x_2(i+1) &= -\exp(-0.016T_i)[0.4\,\sin(0.4T_i)x_1(i) + \sin(0.4T_i-1.531) \\ &\quad x_2(i)] + 0.4\,\exp(-0.016T_i)\sin(0.4T_i)u(i) \end{aligned} \tag{17}$$

where T_i are the switching periods constrained by

$$\sum_{i=1}^{4} T_i \leq 30.5 \quad \text{and} \quad 1 \leq T_i \qquad i = 1, 2, 3, 4 \tag{18}$$

u(i) are the control amplitudes.

Case (i)

Specific control amplitudes are assumed as

$$u(1) = u(3) = 1 \qquad \text{and} \qquad u(2) = u(4) = 2 \tag{19}$$

The desired terminal states are

$$x_1(4) = 2 \quad \text{and} \quad x_2(4) = 0 \tag{20}$$

The state x_1 at switching instants is constrained by

$$x_1(i) \leq 2 \qquad i = 1,2,3 \tag{21}$$

The cost is

$$J = \sum_{i=1}^{3} [x_1^2(i) + x_2^2(i)] \tag{22}$$

Since the system has poor damping and the control amplitudes are fixed, it is to be expected that T_1 is relatively large in order to allow overshoots to die down and $x_1(1) \leq 2$ to be satisfied while keeping $x_1(1)$ and $x_2(1)$ small to minimize the cost.

Case (ii)

The problem is the same as in case (i) except that the control amplitudes are free. The small optimal inputs tend to keep the states at low values until the third sampling period. Large values of u(4) and T_4 are now required to reach the required terminal states. This result is reasonable because the cost does not explicitly contain u(i) and T_i.

Case (iii)

The problem is the same as in case (ii) except that u(i) is bounded by

$$|u(i)| \leq 0.7 \qquad i = 1, \ldots, 4 \tag{23}$$

Now there is a bound on u(i) and the optimal system cannot adopt a large u(4) as in case (ii). The system therefore cannot keep the control and periods in the first three periods at low values as in case (ii). In this case, the optimal system uses the maximum available control amplitude and a large period during the last step.

Case (iv)

In this case the bounds on u are

$$|u(i)| \leq 1.0 \qquad i = 1, \ldots, 4 \tag{24}$$

An energy type cost function

$$J = \sum_{i=1}^{4} [x_1^2(i) + x_2^2(i) + u^2(i)]T_i \tag{25}$$

is to be minimized.

Since $x_1^2(4)T_4$ appears in the cost, and since $x_1(4)$ is constrained to be 2, T_4 has to be reduced. Consequently, the other periods and controls are higher than those in cases (ii) and (iii).

The results of the above 4 cases are tabulated in Table I.

Example 2: The system is a forced Van-der-Pol's equation considered by Merriam (4).

The state equations are

$$\dot{x}_1 = (1-x_2^{\,2})x_1 - x_2 + u$$
$$\dot{x}_2 = x_1 \tag{26}$$

with $x_1(0) = 0$ and $x_2(0) = 1$. The terminal time is fixed at 5 secs and the terminal states are free. A solution with 5 switchings and piecewise-constant control amplitudes is sought to minimize a cost function

$$J = \int_0^5 (x_1^{\,2} + x_2^{\,2} + u^2)dt \tag{27}$$

The problem is to find 4 optimal switching periods and 5 optimal control amplitudes. Discretization with a simple Euler's Formula is used.

Case (i)

There is no constraint on control amplitudes and states. The discrete time results obtained compare favorably with the continuous solutions given by Merriam.

Case(ii)

The control amplitudes are bounded by

$$|u(i)| \leq 0.8 \qquad i = 1,2, \ldots, 5$$

There is a 13% increase in optimal cost over the optimal cost for the unconstrained control of case (i).

Case (iii)

The control amplitudes are constrained by

$$|u(i)| \leq 0.8 \qquad i = 1, 2, \ldots, 5$$

and x_1 is constrained by

$$|x_1(i) \leq 0.35 \qquad i = 1,2$$

The cost is now about 25% more than that for the unconstrained control of case (i). The results are tabulated in Table II and the controls are plotted in fig. (1) for comparison.

V. CONCLUSIONS

The optimal control of discrete-time systems with non-uniform switchings and state control inequality constraints has been solved by using a nonlinear programming technique. The equality constraints are taken care of using a penalty function approach. Examples of linear and nonlinear optimal control problems attempted to date indicate that the method is feasible and may be the only reasonable computational scheme for these type of problems. The method can also be applied to continuous nonlinear systems with nonquadratic type costs provided a sufficiently accurate discretization scheme is employed.

	T_1	T_2	T_3	T_4	u(1)	u(2)	u(3)	u(4)	$T = \sum_{i=1}^{4} T_i$
Case (i)	15.646	1.0	5.795	4.037	1.0	2.0	1.0	2.0	26.49
Case (ii)	2.574	2.982	2.473	7.038	-0.151	-0.224	-0.475	1.06	15.067
Case (iii)	1.006	7.764	4.421	8.823	0.144	0.065	-0.52	0.699	22.014
Case (iv)	17.5	6.93	3.3	2.75	-0.99	-0.81	0.99	-0.301	30.48

Table I. Results of Example 1

	T_1	T_2	T_3	T_4	T_5	u(1)	u(2)	u(3)	u(4)	u(5)	COST
Case (i)	0.527	1.317	0.667	0.637	1.851	0.059	0.958	0.688	0.351	0.037	2.974
Case (ii)	0.9	1.2	0.9	0.5	1.5	0.55	0.8	0.8	0.35	0.05	3.434
Case (iii)	1.122	0.329	1.591	0.738	1.218	0.69	0.8	0.782	0.417	0.093	3.749

Table II. Results of Example 2

REFERENCES

(1) Jordan, B.W., and Polak, E., "Optimal Control of Aperiodic Discrete-Time Systems," J. Siam Control, Ser. A, Vol. 2, No. 3, 1965.

(2) Klingman, W. R., and Himmelblau, D. M., "Nonlinear Programming with the Aid of a Multiple Gradient Summation Technique," Journal of the Association for Computing Machinery, Vol. 11, No. 4, Oct. 1964.

(3) Kuo, B. C., and Tabak, D., "A Nonlinear Programming Approach to Optimal Control of Discrete Systems with Constrained Sampling Times," Preprints of the 9th JACC, University of Michigan, 1968, pp. 918-922.

(4) Merriam, C. W., Optimization Theory and the Design of Feedback Control Systems, McGraw-Hill Book Co., 1964.

APPENDIX

Let $\underline{y}$ be the vector of parameters to be optimized. $f(\underline{y})$ be the cost function, and $C_j(\underline{y})$, j = 1, ... p be the inequality constraints. To start the technique, initial values of $\underline{y}$ and an initial fractional step size Δ have to be supplied. A point $\underline{y}$ which satisfies all constraints is called a feasible point. If the initial point is feasible, optimization begins. If the initial point is not feasible, let

$$C_i(\underline{y}) \quad , \quad i \varepsilon J$$

be the constraints which have negative values. A cost function

$$f'(\underline{y}) = \sum_{i\varepsilon J} C_i(\underline{y})$$

is formulated. Clearly, the maximum value of $f'(\underline{y})$ is zero which is attained when $\underline{y}$ is feasible. The same procedure that is used to minimize the original cost us used to maximize $f'(\underline{y})$ until a feasible point is reached, if the initial is not feasible.

The minimization of $f(\underline{y})$ subject to the inequality constraints $C_j(\underline{y})$, j = 1, 2, ..., p, proceeds as follows. (For a detailed discussion see Klingman & Himmelblau (2)).

(a) Starting from $\underline{y}$, each component y_i of $\underline{y}$ is changed in turn to $y_i \pm \Delta y_i$ and the value which causes a reduction if $f(\underline{y})$ in addition to being feasible is assumed as the new value of y_i. If both $y_i \pm \Delta y_i$ fail to reduce $f(\underline{y})$, the value of y_i is left unchanged. Thus we have a $\underline{y}_{new}$ and a $\underline{y}_{old}$.

(b) A pattern move is made in the above direction with the general rule

$$\underline{y}_{next} = 2\underline{y}_{new} - y_{old}$$

Here again each component is changed in turn so that $\underline{y}_{next}$ is feasible and $f(\underline{y})$ is reduced.

(c) An exploratory move of type II is made following a successful pattern move. This is similar to a move of type I except that the step size used is the same as in move of type I, i.e., the step size = $\Delta\underline{y}_{old}$.

(d) If a pattern move fails or the exploratory move type II following a successful pattern move fails, an exploratory move of type I is performed to define a new direction.

(e) If the move in (d) or (a) fails to find a new point; an adaptive move is made. If $\underline{b}$ is the point at which an adaptive move has to be made, and C_i, i = 1, ...r, are the constraints which will be violated if an exploratory move type I is made at $\underline{b}$, then a new successful direction (NSD) is defined as

$$NSD = \sum_{i=1}^{r} \frac{\text{Grad } C_i}{|\text{Grad } C_i|} + \frac{\text{Grad } f}{|\text{Grad } f|}$$

Let $N = \frac{NSD}{|NSD|}$ where the gradients are evaluated at $\underline{b}$. Let

$N = \sum_i \beta_i \ell_i$ where ℓ_i is the unit vector in the direction of the i^{th} parameter. Let $\alpha_i = \Delta\beta_i$. Now a step is taken from $\underline{b}$ to $\underline{b} + \sum_i \alpha_i \ell_i$. If it is now feasible or if it causes f to increase α_i is systematically reduced as

$$\alpha_i^{n+1} = \alpha_i^{n}/\exp(n)$$

A pattern move is then made in this new successful direction.

(f) If α_i becomes smaller than a predetermined quantity before a successful direction is defined, the process is terminated and a new exploratory move of type I is attempted with a reduced step size Δ.

(g) The whole process is repeated until the change in $\underline{y}$ between two moves is smaller than a preselected amount.

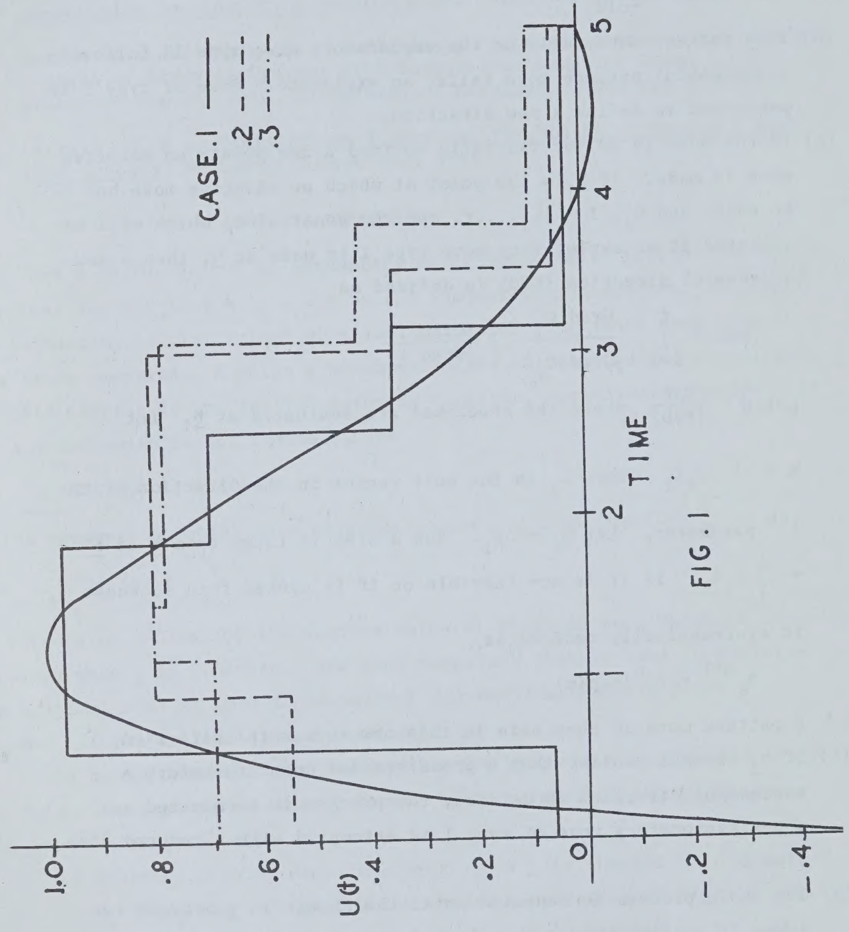


FIG. 1

OPTIMAL LINEAR REGULATORS WITH PRESCRIBED STEADY STATES

MILAN S. ĆALOVIĆ
Department of Electrical Engineering
University of Illinois
Urbana, Illinois

ABSTRACT

This paper considers a problem of regulator design for linear systems with slowly varying measurable disturbances. The design procedure for such regulator systems is developed by introducing the notion of prescribed, disturbance-dependent steady states. In this approach, steady states are not necessarily zero, as they have to be in the standard regulator solution.

This regulator design can be applied to a class of systems in which the cancellation of the effect of disturbances is neither possible nor desirable, by assuming that all states have zero steady-state values. For the modified problem, less restrictive conditions are derived; the range space of the disturbance distribution matrix is not required to be contained in the range space of the control distribution matrix.

As an example, the optimal load and frequency regulator for two interconnected systems is designed so that both optimal transients and desired steady-state characteristics of the system are obtained, satisfying the prescribed steady states.

INTRODUCTION

The basic theory of optimal linear regulators [1,2] was developed for the disturbance-free dynamic systems. The regulators were designed to compensate for the perturbations in the initial conditions, reducing to zero steady state all system states and controls in some optimal fashion. When an input disturbance exists, the standard linear regulators cannot attain and maintain the desired zero equilibrium conditions, and the steady-state error in system states appears.

The need for an extension of the standard linear regulator theory, which will comprise the dynamic systems with constant or slowly varying input disturbances, was emphasized by several authors [3,4,5]. Three main solutions of this problem were proposed:

1. Change of initial condition for the amount of steady-state system error caused by disturbance, and application of the standard regulator theory to the shifted system without disturbance [3].

2. Cancellation of the effect of the disturbance by an additional feed-forward control signal [4].

3. Application of the linear regulator with proportional-plus-integral feedback [5].

The first of these solutions was concerned only with the optimization of transients without affecting the steady-state error, as opposed to the two other approaches considered in the zero steady-state solution. All system states were required to approach zero as time approaches infinity. This requirement cannot be satisfied with a wide class of dynamic systems, where certain range conditions among system matrices are not satisfied (the range space of disturbance distribution matrix F should be contained in the range space of control distribution matrix B). Physically, it means that in such cases there is no control capable of optimizing a given dynamic system resulting in steady states with zero values. The question can be raised, "Do all the system steady states have to be zero in physical systems?" The answer based on considerations of several engineering systems (variable-head hydropower plants, steam boilers, synchronous generating units, power systems, etc.) is negative. In practical applications it is necessary to maintain the zero offsets only at some state variables (named "dominant") while the remaining state variables (named "free") are allowed to take any non-zero steady-state value. State and control variables in steady state are then functions of the disturbance. Within some constraints, these functions can be specified in advance.

This paper considers the problem of linear regulators capable of compensating for known or measurable slowly varying disturbances by permitting system states and controls to have non-zero steady-state values. These steady-state values are prescribed in advance on the basis of physical properties of dynamic systems as linear functions of disturbances.

The theory derived for a general case is then applied to the problem of load and frequency control (LFC) of power systems assuming known load disturbance. The prescribed steady states in this specific problem define the system generation allocation, actually determined by the steady-state system optimization (economic dispatching). Then by also satisfying the other steady-state requirement for zero offsets of frequency and tie-line exchange variations and by providing optimal transients, a linear regulator, as derived in this paper, fulfills all main control objectives in present-day power system control practice.

The paper presents the optimal solution of LFC and simulation results with deterministic load disturbances.

REGULATOR SYSTEM CONFIGURATION

Consider a linear, time-invariant dynamic system with an input disturbance, described by a state-space model

$$\dot{x}(t) = Ax(t) + Bu(t) + Fz(t); \; x(t_o) = x_o, \tag{1}$$

where x, u and z are the state (dimension n), control (dimension m) and disturbance (dimension k) vectors, while A(nxn), B(nxm) and F(nxk) are the constant state, control and disturbance distribution matrices, respectively.

Assuming the disturbance z(t) is constant or a slowly varying quantity, ($z(t) = z \simeq$ constant), the system steady state can be prescribed as a linear function of the disturbance

$$x_{ss} = Mz, \tag{2}$$

where M is the nxk constant matrix. When the matrix $(B'B)^{-1}$ exists, the steady-state system control is

$$u_{ss} = -(B'B)^{-1}B'(AM+F)z. \tag{3}$$

With respect to the relations (2) and (3), system (1) can be reduced to the form of disturbance-free dynamic system

$$\dot{\hat{x}}(t) = A\hat{x}(t) + B\hat{u}(t); \; \hat{x}(t_o) = \hat{x}_o = x_o - x_{ss}, \tag{4}$$

where $\hat{x} = x - x_{ss}$; $\hat{u} = u - u_{ss}$.

The system state vector x can be considered as composed of two components: rxl subvector x^o, which is specified to have zero steady-state value, and (n-r)xl subvector x^f, whose elements are free to take new non-zero steady-state values, which compensate for the effect of the disturbance

$$x = \begin{bmatrix} x^o \\ \hline x^f \end{bmatrix}^* ; \; x_{ss} = \begin{bmatrix} 0 \\ \hline x^f_{ss} \end{bmatrix}. \tag{5}$$

The equation (2) then implies

$$x_{ss} = \begin{bmatrix} 0 \\ \hline M_f \end{bmatrix} z, \tag{6}$$

where M_f is (n-r)xk constant real matrix such that $x^f_{ss} = M_f z$.

The optimal regulator problem for a reduced system model (4) can then be considered as a standard regulator problem when the performance index to be minimized is of the form

$$J = \frac{1}{2}\int_{t_o}^{\infty} (\hat{x}'Q\hat{x} + \hat{u}'R\hat{u})dt, \tag{7}$$

where Q is nxn constant, positive semidefinite symmetric matrix; R is mxm constant, positive definite symmetric matrix.

The control $\hat{u}^*(t)$ that drives the system (4) from an arbitrary initial state $\hat{x}_o$ to zero steady state minimizing the performance index (7), when the system (4) is completely controllable is given by the well-known linear feedback law [1,2,5]

$$\hat{u}^*(t) = -R^{-1}B'P\hat{x}(t), \tag{8}$$

where P is a unique positive definite symmetric solution of the algebraic matrix Riccati equation

$$PA + A'P - PBR^{-1}B'P + Q = 0. \tag{9}$$

In terms of original system (1) variables and matrices, the optimal control (8) is

$$u^*(t) = [R^{-1}B'PM - (B'B)^{-1}B'(AM+F)]\, z - R^{-1}B'Px(t), \tag{10}$$

---

*The elements of the state vector x(t) may always be arranged so that those which are to be with zero steady-state values occupy the first r rows of x(t).

i.e., it represents the standard linear state feedback control, with an additional disturbance and steady-state dependent feed-forward signal.

The closed-loop system

$$\dot{x} = (A-BR^{-1}B'P)x + [F+BR^{-1}PM-B(B'B)^{-1}B'(AM+F)]\, z; \; x(t_o) = x_o, \qquad (11)$$

is asymptotically stable, if the pair [A,D] is completely observable in which D is any matrix such that D'D = Q. The block diagram of this system is given in Fig. 1.

Special Cases

When the matrix M = 0, i.e., when all steady states should be zero, this solution reduces to pure feed-forward disturbance compensation, described in reference [4], where an additional control signal is

$$u_{ss} = - (B'B)^{-1} B'Fz \quad . \qquad (12)$$

Another special case occurs when $M = - A^{-1}F$, i.e., when only optimization of transients is performed, leaving steady states caused by disturbance unaffected, as described in reference [3].

Thus, the solution of the state regulator problem (10) is general. It contains the solutions previously described in references [2,3,4], as special cases.

REGULATOR DESIGN FOR POWER SYSTEM LOAD AND FREQUENCY CONTROL

1. Design Requirements

As an application of the regulator design described in the preceding section, the problem of load and frequency control for two interconnected power systems is considered.

In present day practice LFC is accomplished by using a linear combination of net interchange error and frequency deviation to control the set-points of the system generating units, that is of turbine governors. This control is slowly acting usually of integral type, providing the zero steady-state offsets on frequency and tie-line exchange deviations. It also provides for a steady-state generation power distribution, thus of a system steady-state optimization, with respect to an economic criterion.

Therefore, the conventional LFC is mainly concerned with steady-state system power balance and economy, as opposed to the attempt [3] which employed the standard regulator theory and optimized transients, but it has not satisfied the steady-state features of conventional LFC design. If the features of both conventional and optimal LFC design are required, the objectives of a new design procedure can be stated, assuming that load disturbance is measurable, as

1. zero steady-state offsets of frequency and tie-line variations,
2. optimal transient performance,
3. desired steady-state distribution of generated power.

The modified regulator derived here can be adapted to meet all these requirements; this will be demonstrated in the forthcoming discussion.

2. System Model

A general power system model is developed in reference [6], so that only a short review of the results for a two-system interconnection will be given. The case of power systems composed of one steam, one hydrounit each, is described by the 17-th order model of the form (1), where the system variables and matrices are as follows:

$$x = \begin{bmatrix} x_1 \\ \hline \Delta p_{12} \\ \hline x_2 \end{bmatrix} \quad ; \; u = \begin{bmatrix} u_1 \\ \hline u_2 \end{bmatrix} ; z = \begin{bmatrix} z_1 \\ \hline z_2 \end{bmatrix} ;$$

$$A = \left[\begin{array}{c|c|c} A_1 & a_{12} & 0 \\ \hline m'_{s1} & 0 & -m'_{s1} \\ \hline 0 & -\alpha a_{21} & A_2 \end{array}\right] \quad B = \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \\ \hline 0 & B_2 \end{array}\right] ; \quad F = \left[\begin{array}{c|c} f_1 & 0 \\ \hline 0 & 0 \\ \hline 0 & f_2 \end{array}\right] . \qquad (13)$$

Expressions (13) are given in terms of vectors and matrices for each of the interconnected systems, and intersystem tie-line flow Δp_{12} by using corresponding subscripts 1 or 2 for quantities from one and from the other system respectively. For each of the systems

$$x'_k = [\Delta a_T, \Delta p_{t1}, \Delta p_{t2}, \Delta p_{t3}, \Delta a_H, \Delta v, \Delta q, \Delta f]_k ;$$

$$u'_k = [\Delta f_{mT}, \Delta f_{mH}]_k ; \; z_k = [\Delta p_L]_k .$$

*

$$B'_k = \begin{bmatrix} \frac{1}{T_{sT}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{T_{sH}} & 0 & 0 & 0 \end{bmatrix}_k ; \; a_{12} = f_1 ; \; a_{21} = f_2 ;$$

$$f'_k = [0 \quad 0 \quad 0 \quad 0 \quad \frac{T_a}{T \, T_{sH}} \quad 0 \quad 0 \quad -\frac{1}{T}]_k ; \; k = 1,2. \qquad (14)$$

Intersystem power flow is given by equation

$$\Delta p_{12} = m_s \int_{t_o}^{t} (\Delta f_1 - \Delta f_2) dt = m'_{s1} \int_{t_o}^{t} (x_1 - x_2) dt , \qquad (15)$$

where $m'_{s1} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad m_s]$.

The parameters entering relations (14) are

Δa_T is the variation of steam turbine valve opening, expressed in pu (per unit),

Δp_{t1} is the variation of steam turbine, high-pressure output in pu,

*The matrix for A is typed on the next page.

$$A_k = \begin{bmatrix} -\frac{r_T}{T_{sT}} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & -\frac{1}{T_{sT}} \\ \frac{k_t}{T_u} & -\frac{1}{T_u} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} \\ \text{--} & \frac{1}{T_r} & -\frac{1}{T_r} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} \\ \text{--} & \text{--} & \frac{1}{T_n} & -\frac{1}{T_n} & \text{--} & \text{--} & \text{--} & \text{--} \\ \text{--} & -\frac{\varepsilon_T c_v T_a}{T\,T_{sH}} & -\frac{\varepsilon_T c_s (1-c_v) T_a}{T\,T_{sH}} & -\frac{\varepsilon_T (1-c_v)(1-c_s) T_a}{T\,T_{sH}} & \frac{1}{T_{sH}}\left(\frac{\varepsilon_H k_w T_a}{T} - r_H - r'\right) & \frac{1}{T_{sH}} & -\frac{\varepsilon_H k_q T_a}{T\,T_{sH}} & \frac{eT_a - T}{T\,T_{sH}} \\ \text{--} & \text{--} & \text{--} & \text{--} & \frac{r'}{T_e} & -\frac{1}{T_e} & \text{--} & \text{--} \\ \text{--} & \text{--} & \text{--} & \text{--} & \frac{1}{T_w} & \text{--} & -\frac{1}{T_q} & \frac{1}{T_f} \\ \text{--} & \frac{\varepsilon_T c_v}{T} & \frac{\varepsilon_T c_s (1-c_v)}{T} & \frac{\varepsilon_T (1-c_v)(1-c_s)}{T} & -\frac{\varepsilon_H k_w}{T} & \text{--} & \frac{\varepsilon_H k_q}{T} & -\frac{e}{T} \end{bmatrix}_k$$

Δp_{t2} is the variation of steam turbine, intermediate-pressure output in pu,

Δp_{t3} is the variation of steam turbine, low-pressure output in pu,

Δa_H is the variation of hydroturbine gate opening in pu,

Δv is the variation of dashpot piston position relative to the lever of permanent speed droop of hydroturbine governor in pu,

Δq is the variation of hydroturbine flow in pu,

Δf is system frequency variation in pu,

Δf_{mT} is a steam unit control input variation in pu,

Δf_{mH} is a hydrounit control input variation in pu,

Δp_L is system-load disturbance in pu,

subscripts "T" designate the common parameters relative to steam units,

subscripts "H" designate the common parameters relative to hydrounits,

r is a steady-state speed droop in pu (.05),

T_s is a governor servomotor time constant in sec (T_{sT} = .25; T_{sH} = .10),

T_a is an accelerometric (first time derivative) time constant of hydrogovernor in sec (.8),

T_e is a dashpot time constant in sec (2.0),

T_u is a time constant characterizing the delay in inlet steam piping in sec (.2),

T_r is a time constant of the reheater in sec (6.0),

T_n is a time constant characterizing the delay in low pressure steam piping in sec (.5),

k_t is proportionality factor (.95),

r' is a transient speed droop in pu (.4),

c_v is a fraction of power developed by high-pressure turbine, relative to total unit power output (.3),

c_s is a fraction of power developed by intermediate-pressure turbine, relative to total unit power developed downstream from the reheater (.4),

T_c is a penstock time constant in sec (1.2),

$$T_w = \frac{e_1}{e_2} T_c; \; T_q = e_1 T_c; \; T_f = \frac{e_1}{e_7} T_c; \; k_w = \frac{e_8}{e_1}; \; k_q = \frac{e_5}{e_1}; \; e_H = \frac{e_{10}}{e_1} .$$

e_1 to e_{11} are hydroturbine nondimensional coefficients defined by the following partial derivatives:

$$e_1 = \left(\frac{\partial q}{\partial h_n}\right)_o = .60; \; e_2 = \left(\frac{\partial q}{\partial a}\right)_o = .95; \; e_3 = \left(\frac{\partial \eta}{\partial h_n}\right)_o = .07; \; e_4 = \left(\frac{\partial \eta}{\partial a}\right)_o = .18$$

$$e_5 = \left(\frac{\partial p_t}{\partial h_n}\right)_o = 1 + e_1 + e_3; \; e_6 = \left(\frac{\partial p_t}{\partial a}\right)_o = e_2 + e_4; \; e_7 = \left(\frac{\partial q}{\partial n}\right)_o = 1 - 2_{e1}$$

$$e_8 = e_2 e_5 - e_1 e_7; \; e_9 = \left(\frac{\partial p_t}{\partial n}\right)_o = 1 - 2e_1 - 2e_3; \; e_{10} = e_5 e_7 - e_1 e_9;$$

$$e_{11} = \left(\frac{\partial \eta}{\partial n}\right)_o = - 2e_3;$$

h_n is a net head of hydroplant in pu,

η is a turbine efficiency in pu,

n is a speed of hydroturbine rotation in pu,

p_t is a hydroturbine output in pu,

T is a system acceleration time constant in sec (12),

$e = e_p - \varepsilon_T e_T + \varepsilon_H e_H$ is system self-regulation coefficient in pu,

e_p is a consumer self-regulation coefficient in pu (1.5),

e_T is a steam turbine self-regulation coefficient in pu (.15),

ε is a participation coefficient of particular unit in total system load ($\varepsilon_{T1} = .4$; $\varepsilon_{H1} = .5$; $\varepsilon_{T2} = .5$; $\varepsilon_{H2} = .4$; $\varepsilon_L = .1$; subscript "L" designates the units operating with disconnected speed-sensitive elements at turbine governors).

$\alpha = P_{10}/P_{20}$ is the size ratio coefficient, i.e., the ratio of steady-state loads of systems 1 and 2 (1.0),

$m_s = k.2\pi F_o (\cos\theta_{12})_0$ is the synchronizing power flow coefficient expressed in sec^{-1},

$k = \dfrac{P_{12m}}{P_{10}}$ is tie-line capacity, relative to the steady-state load of system 1, in pu (.1),

$(\theta_{12})_0 = \theta_{10} - \theta_{20}$ is initial phase-angle difference between two systems, in rad. ($\pi/4$).

The numerical values appearing in parenthesis represent the basic system parameters used in optimization and simulation studies.

3. System Specifications

After defining the basic system model (1) with numerical values of the elements in the expressions (13) to (15), the specification of other quantities appearing in analytic expressions (1) to (11) should be given. These are

a. state subvector x^o, desired to have zero steady-state value,
b. matrix M, which defines desired steady states,
c. penalty matrices Q and R.

Subvector x^o represents the frequency and load-exchange variations, the variables whose steady states must be zero by design requirement. Thus,

$$x'^o = [\Delta f_1, \Delta p_{12}, \Delta f_2]; \; r = 3. \tag{16}$$

Matrix M is selected by solving the system of n algebraic equations (z is assumed to be known constant)

$$Ax_{ss} + Bu_{ss} + Fz = 0. \tag{17}$$

In general, the system (17) of n algebraic equations with (n+m) unknowns is consistent (i.e., possesses solutions) if and only if the rank of the coefficient matrix [A,B] equals the rank of the augmented matrix [A,B,-Fz]. Whenever the solution exists, all solutions can be expressed in terms of m independent parameters, providing that matrix A has full rank n. By adding to the system (17) $(r-\ell)$ equations $(r-\ell \leq m)$ from the condition $x^o_{ss} = 0$, the solutions of (17) will be given in terms of $(m-r+\ell)$ independent parameters, where ℓ is the number of homogeneous equations in (17), which reduce to the relationships among the components of x^o_{ss}. For the system model considered here, $\ell = 1$, so that $m-r+\ell = 2$, i.e., the solution of (17) will be given in terms of two independent parameters. These parameters are chosen to be the steady-state values of system controls w_{1ss} and w_{2ss}, such that

$$u_{ss} = \nu w_{ss}, \tag{18}$$

where ν is the block-diagonal matrix composed of vectors, the elements of which define the steady-state participation of each generating unit in LFC. They are determined by an optimization of the steady-state distribution of generated power. Thus,

$$\nu_k = \begin{bmatrix} \nu_T \\ \nu_H \end{bmatrix}_k ; \; k = 1,2; \; \nu_T + \nu_H = 1.00. \tag{19}$$

The solution of (17), with respect to (18) and (19) defines the matrix M as

$$M = \left[\begin{array}{c:c} M_1 & 0 \\ \hdashline 0 & 0 \\ \hdashline 0 & M_2 \end{array}\right], \tag{20}$$

where $M'_k = \dfrac{1}{a_k}\left[\dfrac{\nu_T}{r_T}, \dfrac{\nu_T k_t}{r_T}, \dfrac{\nu_T k_t}{r_T}, \dfrac{\nu_T k_t}{r_T}, \dfrac{\nu_H}{r_H}, \dfrac{\nu_H r'}{r_H}, \dfrac{\nu_H e_2}{r_H}, 0\right]_k$;

$$a_k = \left[\frac{\varepsilon_T \nu_T k_T}{r_T} + \frac{\varepsilon_H \nu_H e_6}{r_H}\right]_k ; \; k = 1,2.$$

The selection of $\nu_1 = \begin{bmatrix} .4 \\ .6 \end{bmatrix}$; $\nu_2 = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$, yields

$$M'_1 = \frac{1}{9.82}\,[8.0 \quad 7.6 \quad 7.6 \quad 12.0 \quad 4.8 \quad 11.4 \quad 0]$$

$$M'_2 = \frac{1}{9.27}\,[10.0 \quad 9.5 \quad 9.5 \quad 10.0 \quad 4.0 \quad 9.5 \quad 0]. \tag{21}$$

The matrices R and Q entering the performance index (7) are selected so that a 4 x 4 matrix R = I, while the 17 x 17 matrix Q has all zero elements except that r = 3 diagonal terms; this corresponds to the penalization of x^o. Two values of the non-zero terms of matrix Q are used: 1 and 100. The LFC regulator design is executed for two equal systems, i.e., for a size-ratio coefficient $\alpha = 1$.

5. Computer Results

A FORTRAN IV program for a given system (1) and set of matrices ν, M, Q, and R has the following steps:

1° Determination of open-loop system eigenvalues.
2° Check of controllability and observability.
3° Solution of Riccati equation (9).
4° Calculation of optimal feedback gains (10).
5° Calculation of closed-loop system (11) eigenvalues.

6° Optimum system (11) response simulation assuming step-load disturbance.

7° Plotting of response curves obtained in step 7° (Calcomp plots).

For the solution of the algebraic matrix Riccati equation (9), the eigenvectors method [7] is used.

All of the calculations were performed on the digital computer IBM 360/75 of the University of Illinois. A typical figure for complete time needed to get an optimal solution with simulation (steps 3° to 6°) was about 55 seconds, with memory usage of about 150k bytes.

When considering the feedback gains, it can be noted that optimal LFC is much more influenced by the states from its own power system, with less significant contribution of the states from the neighboring system.

The check of closed-loop system eigenvalues shows an important increase of stability margin when optimal control is compared with conventional control design [6].

Figure 2 shows the curves of frequency and tie-line variations for a free-system response upon 1% step-load change in system 1, while Fig. 3 shows the same variables when conventional LFC is applied.

System response under optimal control is illustrated in Fig. 4 for the same disturbance conditions as in Figs. 2 and 3.

The analysis of these curves shows the considerable improvement in system transients when optimal control is applied for the same steady-state performance as with the conventional LFC. An especially favorable effect of the optimal control is in damping of fast transients of tie-line load variations, inherently very oscillatory. A higher ratio of the coefficients of penalty matrices Q and R gives higher feedback gains and faster rates of decay of response curves reducing at the same time the overshoot peaks (increased damping).

CONCLUSIONS

The design procedure for optimal linear regulators in the dynamic systems with slowly varying measurable disturbances, by specifying in advance the system steady states as linear function of disturbances, is presented. It is shown in the example of load and frequency control in two-power system interconnection that the proposed design procedure provides desired steady-state features by greatly improving system transients and stability margins.

ACKNOWLEDGMENT

The author wishes to express his appreciation to the ALCOA Foundation, to the Departments of Electrical Engineering, Mechanical Engineering, and the Coordinated Science Laboratory of the University of Illinois, Urbana, Illinois, for their support in this research. He is also grateful to Professors J. B. Cruz, Jr., W. R. Perkins, and P. V. Kokotović for their interest and suggestions given during the preparation of this work. In addition the assistance of Dr. Gurdial Singh in computer programming is acknowledged and his valuable discussions throughout this work are greatly appreciated.

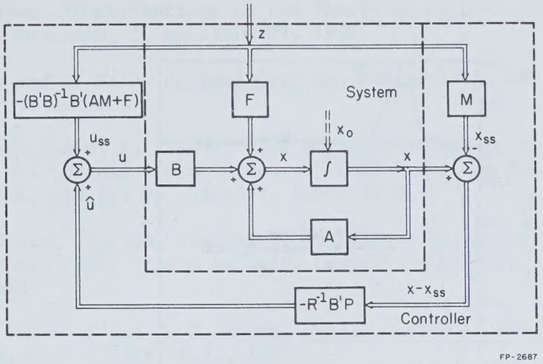
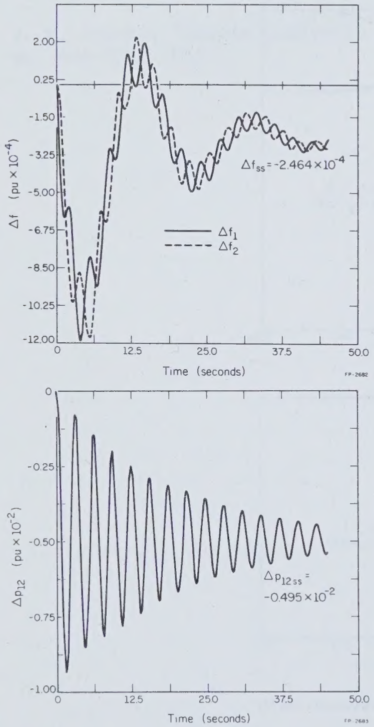

z
-(B'B)$^{-1}$B'(AM+F)
F
System
M
u$_{ss}$
x$_0$
x$_{ss}$
u
x
x
B
Σ
∫
Σ
Σ
û
A
-R^{-1}B'P
x-x$_{ss}$
Controller
FP-2687

Fig. 1. Block diagram of linear regulator with prescribed steady states.



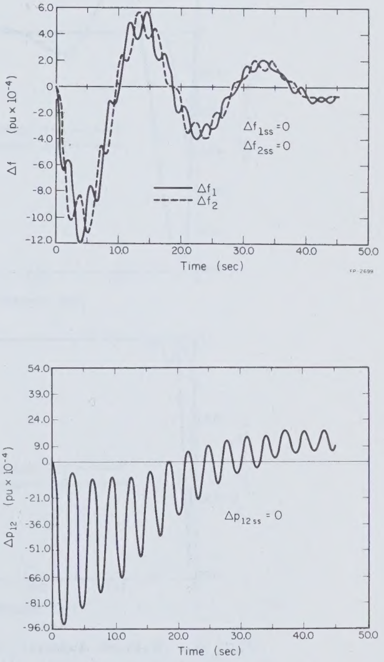


Fig. 2. Free system response curves. Fig. 3. System response with conventional LFC.

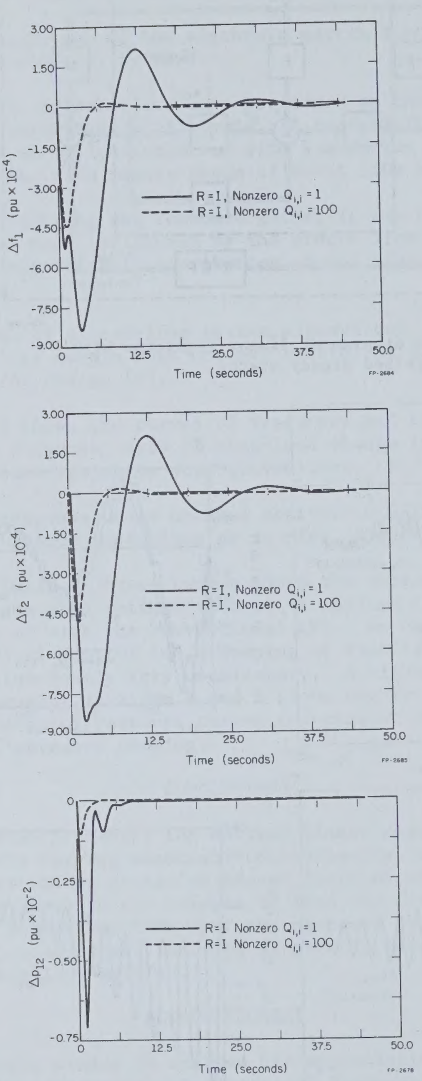


Fig. 4. System dynamic response with optimal LFC

REFERENCES

1. R. E. Kalman, "Contributions to the Theory of Optimal Control," Bol. Soc. Mat. Mexicana, 5, pp. 102-109, 1960.

2. M. Athans and P. Falb, Optimal Control, McGraw-Hill, New York, N.Y., 1966.

3. C. E. Fosha and O. E. Elgerd, "The Megawatt-Frequency Control-A New Approach via Optimal Control Tneory," IEEE Trans. on Power Apparatus and Systems, PAS 89, pp. 563-571, April 1970.

4. C. D. Johnson, "Optimal Control of the Linear Regulator with Constant Disturbances," IEEE Trans. on Auto. Control, AC-13, pp. 416-421, August 1968.

5. B. D. O. Anderson and J. B. Moore, Linear Optimal Control, Prentice-Hall, Englewood Cliffs, N.J., 1971.

6. M. Calović, "Dynamic State-Space Models of Electric Power Systems," University of Illinois, Dept. of EE Report, Urbana, Illinois, 1971.

7. J. E. Potter, "Matrix Quadratic Solutions," J. SIAM Apl. Math., pp. 496-501, 1966.

A GRADIENT METHOD FOR THE OPTIMAL CONTROL PROBLEM OF BOLZA

SHYH JONG WANG
GTE Laboratories, Inc.
Waltham, Massachusetts

EDGAR C. TACKER
Louisiana State University
Baton Rouge, Louisiana

ABSTRACT

An algorithm for the iterative solution of the optimal control problem of Bolza with isoperimetric and algebraic constraints is introduced. A physical problem is considered and results are presented.

INTRODUCTION

Over the past several years a number of computational methods have been developed for solving the optimal control problem of Mayer. In order to apply these methods to the problem of Bolza one has to transform the Bolza problem into the Mayer problem by introducing additional state and adjoint variables. This often results in increasing memory requirements and computing time, unless special treatment is given beforehand. The purpose of this paper is to overcome this difficulty by introducing a new algorithm which is a generalization of earlier gradient methods [1,2]. The new formulation facilitates a uniform treatment of the problems of Bolza, Mayer and Lagrange.

PROBLEM STATEMENT

The optimal control problem can be formulated as the following [3]:

$$J = g(t_f,x_f,a) + \int_{t_o}^{t_f} L(t,x,u,a)dt \qquad (1)$$

$$\dot{x} = f(t,x,u,a) \qquad (2)$$

$$I^\gamma \le 0,\ \gamma = 1,..,p';\ I^\gamma = 0,\ \gamma = p'+1,...,p \qquad (3)$$

$$I^\gamma = \theta^\gamma\ (t_f,x_f,a) + \int_{t_o}^{t_f} M^\gamma(t,x,u,a)dt \qquad (4)$$

where x, u, a are the n-state, m-control and r-control parameter vectors, respectively. J is the cost functional and I is the p-isoperimetric constraint. For computational purposes a stopping functional S is defined as

$$S(t_f) = \theta^o(t_f,x_f,a) + \int_{t_o}^{t_f} M^o(t,x,u,a)dt \qquad (5)$$

where S = 0 defines the terminal time t_f. It is assumed that the (total) derivative of S at t_f is nonvanishing. The algebraic constraints are:

$$\phi^s(t,x,u,a) \le 0,\quad s = 1,...,q';\ \phi^s(t,x,u,a) = 0,\ s = q'+1,...,q \qquad (6)$$

The inequality constraints (3) are transformed into equality constraints by introducing p′ addional control parameters β^γ and the algebraic constraints are treated by adding weighted penalty functions [4,5] to the integrand of J. The control problem of Bolza is to find the control u(t),a,β and the arc x such that J is minimized and (2), (3) and (6) are satisfied.

THE ALGORITHM

The algorithm is developed by considering the first-order effect on the system introduced by applying a small perturbation along the nominal solution arc. Influence functions λ_J, λ_S and λ_I are used to generate the first variations of J,S and I and the optimal perturbations in u,a,β,etc., are obtained under the constraint

$$\int_{t_o}^{t_f} \delta u^T\ U\ \delta u\ dt + da^T\ A\ da + d\beta^T\ W\ d\beta + db^T\ B\ db - (dC)^2 = 0$$

where U(t), A, W and B are mxm, rxr, p′xp′ and (n+1)x(n+1) symmetric positive definite weighting matrices, respectively. (dC) is a scalar whose value, in general, decreases from iteration to iteration under program control and $db^T = \{dt_o\ dx^T(t_o)\}$. The details of the algorithm are given in reference [6]. The computation starts from a nominal solution u^*, a^*, β^*, b^*. Equation (2) is then integrated forward from t_o until S=0 at t_f and then the adjoint equations are integrated backward from t_f to t_o with $\lambda_J^T(t_f) = \partial g/\partial x_f$, $\lambda_S^T(t_f) = \partial\theta^o/\partial x_f$ and $\lambda_I^T(t_f) = \partial\theta/\partial x_f$ as boundary conditions. The optimal "steps" $\delta u(t)$, da, $d\beta$, and db are then computed and added to $u^*(t)$, a^*, β^*, b^* to obtain a new nominal solution. In each succeeding iteration, actual and predicted dJ and dI are computed, and the relative prediction errors are used for guidance in choosing (dC), dI and the weighting matrices for the next iteration. If the errors are excessively large it is often best to return to the previous iteration and iterate the computation with different (dC), etc., and hence both the old and the new controls along with new state x(t) have to be stored in memory. The iterative process stops when a desired solution is approached or when (dC) reaches a preselected small value.

NUMERICAL EXAMPLE--A TRAJECTORY OPTIMIZATION PROBLEM--RENDEZVOUS

This example concerns the trajectory optimization problem of Earth-Venus orbital rendezvous, with fixed final time and vehicle heliocentric angle, of a low-thrust rocket with variable thrust and thrust angle. The normalized dynamic equations, the cost functional, the isoperimetric constraints and the stopping function are, respectively,

$$\dot{x}^1 = 10^{-6}(x^2)^2/x^3 + 10^{-3}u^2/x^4 - .13249(x^3)^{-2} = f^1 \qquad x^1(0) = 0$$

$$\dot{x}^2 = -10^{-6}x^1x^2/x^3 + 10^{-3}u^1/x^4 = f^2 \quad \text{N=Iteration number} \qquad x^2(0) = 29.76$$

$$\dot{x}^3 = 10^{-6}x^1 = f^3 \qquad x^3(0) = 149.6$$

$$\dot{x}^4 = -1.7918(10^{-5})[(u^1)^2 + (u^2)^2]^{\frac{1}{2}} = f^4 \qquad x^4(0) = 679.602$$

$$J = \int_o^{t_f} u^T \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix} u\, dt, \quad w_1 = w_2 = .00002$$

$$I^1 = \int_o^{t_f} 10^{-6}\, x^2/x^3 dt - \phi_f = 0 \quad (\phi_f = 4\pi/3), I^2 = x^1(t_f) = 0$$

$$I^3 = x^2(t_f) - 35 = 0, \qquad I^4 = x^3(t_f) - 108.2 = 0$$

$$S(t_f) = t_f - 17{,}280{,}000 = 0 \quad (200 \text{ days})$$

Nominal control , N=0 :

$u^1(t)$	$u^2(t)$	duration (sec)
-.556	.0981	2,592,000
-.363	.4325	4,320,000
.0869	.5620	10,368,000

where u^1 and u^2 are the thrust components in the circumferential and radial directions, respectively. In this particular example, $\lambda_J, \lambda_S = 0$, the effect of J is introduced into the algorithm through the Hamiltonian H_{JS}. In each iteration, 6 equations are integrated forward and 37 equations are integrated backward. The following $(N, J, I^1, I^2, I^3, I^4)$ data (rounded) illustrates the effectiveness of our algorithm: (0,111.1313,-.1809,2.3174,.2791, 4.1862),(4,59.2103,-.0784,.9616,.0814,1.899),(30,23.3088,-.0004,.0002, .0008,.0016), and (38,23.2956,-.0001,-.0003,.0002,.0002).

REFERENCES

[1] Kelley, H.J., "Gradient Theory of Optimal Flight Paths", ARS Journal 30, pp.947-954, Oct. 1960.

[2] Bryson, A.E.; Denham, W.F., "A Steepest-Ascent Method for Solving Optimum Programming Problems", J. Of Applied Mechanics, pp.247-251,1962.

[3] Hestenes, M.R., "Calculus of Variations and Optimal Control Theory", John Wiley & Sons, Inc., 1966.

[4] Okamura, Kiyohisa, "Some Mathematical Theory of the Penalty Method for Solving Optimal Control Problems", J. SIAM Control, Ser. A,V2,N3,1965.

[5] Lele M.M. & Jacobson, D.H., "A Proof of the Convergence of the Kelley-Bryson Penalty Function Technique for State-Constrained Control Problems", J. MAA, Vol. 26, pp. 163-169, April 1969.

[6] Wang, S.J., "A Gradient Computational Technique for a Class of Optimal Control Problems Subject to Inequality Constraints", Ph.D. thesis,1969 Michigan State University.

NECESSARY CONDITIONS OF OPTIMALITY-A UNIFIED APPROACH

VIJAY P. BHATKAR AND V.S. RAJAMANI
Department of Electrical Engineering
Indian Institute of Technology, New Delhi-29, India

ABSTRACT

If generalized functions are admitted, a wide class of optimal control problems can be transcribed into a canonical form called the Canonical Variational Problem (CVP). The necessary optimality condition for the CVP is phrased in the form of Lagrange multiplier rule. This condition contains most of the results relating to necessary conditions in the theory of optimal control.

INTRODUCTION

During the last decade considerable research has been devoted to extending the original minimum principle of Pontraygin [1] - - which gives a set of necessary optimality conditions for differential systems — to include differential-difference, partial differential, integral, integro-differential, and other complex systems (see, for example [2]-[6]; numerous additional references can be found in the survey articles [7], [8]). It is now possible to unify many of these results within a single framework, and it is to this that the present paper is devoted. The central idea of unification is based on the observation that, if generalized functions are admitted, a wide class of optimal control problems can be transcribed into a canonical form which is a generalization of the isoperimetric problem in the calculus of variations with several independent variables. This canonical form is christened as the Canonical Variational Problem (CVP). The necessary optimality condition for the CVP is phrased in the form of Lagrange multiplier theorem. Most of the known necessary conditions of optimality can be obtained as direct and straightforward consequences of this fairly simple, but unifying theorem.

The notation is, in most respects, adopted from Edelen [9]. The summation convention is assumed throughout. Thus, if any indexed quantity has the same index letter twice, that index is to be summed over the range of the index involved.

CANONICAL VARIATIONAL PROBLEM

Let D be an open, arc-wise simply connected point set in n-dimensional number space X_n and let D^* denote the closure of D with respect to X_n. Let (x) be a coordinate system in D^*. The "coordinate differential volume element" of D will be denoted by dV(x). The domains of definition of (x) and (z) will be assumed to be the same.

Let $\varphi_\Lambda(x^k)$, $\Lambda = 1,\dots,N$, denote an ordered collection of N functions on D^*. We shall refer to $\varphi_\Lambda(x^k)$ as a set of manipulatable variables defined on D^*. Let Ω denote a compact subset of X_n.

Define the Lagrangian function L on $D^* \times \Omega$, g_a, $a = 1,\dots,q$, on $d^*_m \times D^* \times \Omega$ and let

$$J[\varphi_\Lambda] = \int_{D^*} L(x^k, \varphi_\Lambda(x^k))\,dV(x) \qquad (1)$$

$$k_a(x^b;\varphi_\Lambda) = \int_{D^*} g_a(x^k, z^k, \varphi_\Lambda(z^k))\,dV(z) \equiv c_a(x^k) \qquad (2)$$

where d^*_m denotes closed subset of D^*.

The problem is to minimize $J[\varphi_\Lambda]$ over $\varphi_\Lambda \varepsilon\, \Omega$ subject to the generalized isoperimetric constraint (2). We refer to this problem as the Canonical Variational Problem.

Suppose that L is continuous on $D^* \times \Omega$, and absolutely integrable on D^* for each $\varphi_\Lambda(x^k) \varepsilon\, L^1(D^*;N)$— a generic notation for N-tuple functions absolutely integrable over D^*. Similarly, suppose that g_a are continuous on $d^*_m \times D^* \times \Omega$ and absolutely integrable on D^* for each $\varphi_\Lambda(x^k) \varepsilon\, L^1(D^*;N)$. We do not make any assumption about the structure of the set Ω and the smoothness character of the functions L and g_a.

LAGRANGE MULTIPLIER THEOREM

Clearly, the CVP falls outside the classical calculus of variations: it cannot be solved by the direct variation of $\varphi_\Lambda(x^k)$. However, devices exist which enable us to characterize its solution. One of them consists of using the method of introducing sliding regime [10] in which the variations of $\varphi_\Lambda(x^k)$ are replaced by the variations of some auxiliary function, thereby avoiding any assumptions about the structure of the set Ω. By extending the methods given in [10] and [11], the minimum condition for the CVP can be derived. We shall, for reason of brevity, omit the proof and simply spell out the final result.

THEOREM. If an optimal solution φ^0_Λ exists, then there exists a non-zero q-tuple Lagrange multiplier function λ^a such that φ^0_Λ minimizes the functional

$$\bar{J}[\varphi_\Lambda, \lambda^a] = J[\varphi_\Lambda] + \int_{d^*_m} \lambda^a(x^k)\,k_a(x^k;\varphi_\Lambda)\,dV(x) \qquad (3)$$

with respect to all $\varphi_\Lambda \varepsilon \hat{\Omega}$, where $\hat{\Omega}$ is the set of permissible functions φ_Λ. Further, if $\lambda^a g_a$ is a summable function when $x^k \varepsilon d^*_m$, $z^k \varepsilon D^*$ then the minimum condition

$$\mathcal{L}(x^k,\varphi^0_\Lambda(x^k),\lambda^a(x^k)) = \min_{\varphi_\Lambda \varepsilon \Omega} \mathcal{L}(x^k,\Phi_\Lambda(x^k),\lambda^a(x^k)) \qquad (4)$$

holds at each x^k, where $\mathcal{L}$ is the extended Lagrangian function defined by

$$\mathcal{L} = L(x^k,\varphi_\Lambda(x^k)) + \int_{d^*_m} \lambda^a(z^k) g^*_a dV(z) \qquad (5)$$

$$(g^*_a = g_a(z^k,x^k,\varphi_\Lambda(x^k))). \qquad (6)$$

A salient feature of this theorem is that it is not valid with respect to arbitrary variations about φ^0_Λ. This is due to the fact that the use of the Lagrange multipliers is valid only for those variations about φ^0_Λ for which the increment in the functional J and each of the functional transformations k_a is linearly dependent on them. Clearly, such variations — and hence the structure of the set of permissible functions $\hat{\Omega}$ — are governed by the form of L and g_a. The problem of determining the structure of $\hat{\Omega}$ from a knowledge of L and g_a will be treated elsewhere. Transcribing various types of constraints into the form (2), one obtains g_a and $\mathcal{L}$ for each type of constraint, g_a dictates the set of permissible functions (for each component of φ_Λ), and the minimum condition of $\mathcal{L}$ over the set of permissble functions is the required optimality condition.

If the variable $X_j = \varphi_j$ occurs in the constraint equation in such a way that only infinitesimally small variations are permissible with respect to it at each x^k, then the minimum condition takes the form

$$(\partial/\partial X^0_j)\mathcal{L} = 0 \qquad (7)$$

provided L and g_a are of class C^1 in X_j; and the minimum is reached within the set of permissible values of X_j. In optimal control theory such variables are usually the state variables of the system. If finite variations are permissible with respect to the variable $U_i = \varphi_i(x^k)$ and U_i is constrained by the condition $U_i \varepsilon \Omega$ then, with respect to that component, the minimum condition takes the form

$$H[U^0_i] = \min_{U_i \varepsilon \Omega} H[U_i]. \qquad (8)$$

Here H is the Hamiltonian function and it makes up that part of $\mathcal{L}$ which involves U_i explicitly. When H is of class

C^1 in U_i and Ω is open, then the condition

$$(\partial/\partial U_i^0)H = 0 \tag{9}$$

can also be used for this variable. In the context of optimal control theory such variables are usually the control variables. If, however, Ω includes all possible values of φ_Λ, then the necessary minimum condition is also sufficient.

EXAMPLES

We now illustrate how the theorem works by means of two examples.

Coupled ordinary and partial differential system[3]. Consider the problem of minimizing

$$J[U,X,Z,W,Y] = \int_0^T L_1(t,U(t),X(t),Z(t))\,dt + \int_0^T \int_0^L L_2(t,x,W(t,x),Y(t,x))\,dt\,dx \tag{10}$$

subject to the coupled ordinary and partial differential constraint:

$$\dot{X}(t) = f(t,U(t),X(t),Z(t)),\quad t \in [0,T] \tag{11-a}$$

$$\partial_t Y(t,x) + g(t,x)\partial_x Y(t,x) = F(t,x,W(t,x),Y(t,x)),\quad t \in [0,T],\ x \in [0,L] \tag{11-b}$$

$$X(0) = X_0;\ Y(0,x) = Y_0(x),\ x \in [0,L];\ Y(t,0) = X(t),\quad t \in [0,T] \tag{11-c}$$

$$Z(t) = Y(t,L),\quad t \in [0,T]. \tag{11-d}$$

We first transcribe the problem into the CVP:

$$J[U,X,Z,W,Y] = \int_0^T \int_0^L \left(\frac{1}{L} L_1 + L_2\right) dt\,dx \tag{12}$$

$$k_1(t;U,X,Z) = \int_0^T \int_0^L \frac{1}{L}[-X(\tau)\delta(t;\tau) + f^* h(t;\tau)]\,d\tau\,dz \equiv -x_0 \tag{13-a}$$

$$\begin{aligned} k_2(t;x;X,W,Y) = \int_0^T \int_0^L [&-Y(\tau,z)\delta(t;\tau)h(x;z) - g(\tau,z)Y(\tau,z) \\ &h(t;\tau)\delta(x;z) + \frac{1}{L} g(\tau,0)X(\tau)h(t;\tau) + \\ &\partial_z g(\tau,z)h(t;\tau)h(x;z) + F^* h(t;\tau)h(x;z)]\,d\tau\,dz \\ \equiv -\int_0^x &Y_0(z)\,dz \end{aligned} \tag{13-b}$$

$$k_3(t;X,Z,Y) = \int_0^T \int_0^L [-Y(\tau,z)\delta(t;z) - \frac{1}{L}g(\tau,L)Z(\tau)h(t;\tau) + \frac{1}{L}g(\tau,0)X(\tau)h(t;\tau) + \partial_z g(\tau,z)Y(\tau,z)h(t;\tau) + F^* h(t;\tau)]d\tau dz \equiv -\int_0^L Y_0(z)\,dz \qquad (13\text{-c})$$

where $\delta(\cdot)$ denotes the Dirac delta function, $H(\cdot)$ denotes the Heavyside unit step function, and the symbol $*$ has the same meaning as in (6). In this case, from (5), we have the extended Lagrangian

$$\begin{aligned} \mathcal{L} = {} & \frac{1}{L} L_1 + L_2 - \lambda^1(t)X(t) + f\int_t^T \lambda^1(\tau)\,d\tau - Y(t,x)\int_x^L \lambda^2(t,z)\,dz \\ & - g(t,x)Y(t,x)\int_t^T \lambda^2(\tau,x)\,d\tau + \frac{1}{L}g(t,0)X(t)\int_t^T \int_0^L \lambda^2(\tau,z)\,d\tau\,dz \\ & + \partial_x g(t,x)Y(t,x)\int_t^T \int_x^L \lambda^2(\tau,z)\,d\tau dz + F\int_t^T \int_x^L \lambda^2(\tau,z)\,d\tau\,dz \\ & - \lambda^3(t)Y(t,x) + \frac{1}{L}g(t,L)Z(t)\int_t^T \lambda^3(\tau)\,d\tau \\ & + \frac{1}{L}g(t,0)X(t)\int_t^T \lambda^3(\tau)\,d\tau) + \partial_x Y(t,x)\int_t^T \lambda^3(\tau)\,d\tau \\ & + F\int_t^T \lambda^3(\tau)\,d\tau. \end{aligned} \qquad (14)$$

We introduce the adjoint variables

$$p(t) = \int_t^T \lambda^1(\tau)\,d\tau; \quad q(t,x) = \int_t^T \int_x^L \lambda^2(\tau,z)\,d\tau\,dz + \int_t^T \lambda^3(\tau)\,d\tau \qquad (15)$$

and the Hamiltonian functions

$$H = L_1 + p(t)f; \quad M = L_2 + q(t,x)F. \qquad (16)$$

The conditions $(\partial/\partial X^0)\mathcal{L} = 0$, $(\partial/\partial Y^0)\mathcal{L} = 0$, and $(\partial/\partial Z^0)\mathcal{L} = 0$ then yield, after some simplification, the following adjoint system:

$$p(t) = -(\partial/\partial X^0)H - g(t,0)q(t,0) \tag{17-a}$$

$$\partial_t q(t,x) + \partial_x (g(t,x)q(t,x)) = -(\partial/\partial Y^0)M \tag{17-b}$$

$$(\partial/\partial Z^0)H = q(t,L)g(t,L) \tag{17-c}$$

$$p(T) = 0; \quad q(T,x) = 0 \tag{17-d}$$

Note that (17-d) is a consequence of (15). The minimum condition requires the control variables U(t) and W(t,x) to be chosen so that the functions H and M attain their minimum values.

The adjoint system and the minimum condition are exactly the same as those obtained by Hiratsuka and Ichikawa [3]. It is worthwhile to remark that the system (11) embeds a large class of differential difference systems, and hence the preceding results serve to abstract most of the results reported in [1], [2].

Integrodifferential Systems. Next, consider the problem of minimizing

$$J[U,X] = \int_0^T \int_0^L L(t,x)\ U(t,x), X(t,x), \partial_t X(t,x), \partial_x X(t,x)) \, dt\, dx \tag{18}$$

subject to the integrodifferential constraint:

$$\partial_t \partial_x X(t,x) = \int_0^T \int_0^L g(t,x,\tau,z,U(\tau,z),X(\tau,z),\partial_\tau X(\tau,z), \partial_z(\tau,z)) d\tau \, dz \tag{19-a}$$

$$X(0,x) = \alpha(x); \; X(t,0) = \beta(t); \; \alpha(0) = \beta(0) \tag{19-b}$$

The system (19) is an importnt generalization of the Goursat system considered by Egorov [4]. We set

$$Y(t,x) = \partial_t X(t,x), \quad Z(t,x) = \partial_x(t,x) \tag{19-c}$$

Before we transcribe the above constraint into the required form, let us recall certains useful properties of multi-dimensional delta functions. Specifically, for our immediate purpose, we note:

$$\int_d D^p\delta(x^k;z^k)\;\varphi(z^k)\,dV(z) \;=\; D^p\varphi(x^k) \tag{20}$$

$$\int_d D^p\delta(z^k;x^k)\varphi(z^k)\,dV(z) = (-1)^{|p|}\varphi(x^k) \tag{21}$$

where

$$D^p = \prod_m(\partial_m)^{p_m} \quad ; \quad |p| = \sum_m p_m \tag{22}$$

and $\varphi(x^k)$ is a test function [10] on d. Using (20), we write (19) as

$$k_1(t,x;U,X,Y,Z) = \int_0^T\int_0^L [-X(\tau,z)\partial_\tau\partial_z\delta(t,x;\tau,z)+g]\,d\tau dz \equiv 0 \tag{23-a}$$

$$k_2(t,x;X,Y) = \int_0^T\int_0^L [-X(\tau,z)\,\delta(t,x;\tau,z)+Y(\tau,z)h(t;\tau)\,\delta(x;z)]\,d\tau dz \equiv -\alpha(x) \tag{23-b}$$

$$k_3(t,x;X,Z) = \int_0^T\int_0^L [-X(\tau,z)\,\delta(t,x;\tau,z)+Z(\tau,z)\,\delta(t;\tau)h(x;z)]\,d\tau\, dz \equiv -\beta(t). \tag{23-c}$$

We then have

$$\begin{aligned} \mathcal{L} = {} & L-\partial_t\partial_x\lambda^1(t,x)X(t,x)+\int_0^T\int_0^L \lambda^1(\tau,z)g^*\,d\tau\,dz \\ & -\lambda^2(t,x)X(t,x)+Y(t,x)\int_t^T \lambda^2(\tau,x)\,d\tau \\ & -\lambda^3(t,x)X(t,x)+Z(t,x)\int_x^L \lambda^3(t,z)\,dz \end{aligned} \tag{24}$$

$$H = L + \int_0^T\int_0^L \lambda^1(\tau,z)g^*\,d\tau dz \tag{25}$$

The conditions $(\partial/\partial X^0)\mathcal{L} = 0$, $(\partial/\partial Y^0)\mathcal{L} = 0$, $(\partial/\partial Z^0)\mathcal{L} = 0$ yield the adjoint system:

$$\partial_t \partial_x \lambda^1(t,x) = (\partial/\partial X^0)H - \partial_t((\partial/\partial(\partial_t X^0))H) - \partial_x((\partial/\partial(\partial_x X^0))H) \qquad (26\text{-a})$$

$$(\partial/\partial(\partial_t X^0))H|_{t=T} = (\partial/\partial(\partial_x X^0))H|_{x=L} = 0, \qquad (26\text{-b})$$

Observe that the integral systems of Butkovskii [5], the Goursat systems of Egorov [4], the integrodifferential systems of Bate [6] — in fact, many of the system models used in the literature of distributed systems— are just the special cases of (19), and hence their results are immediate consequences of the above results.

CONCLUDING REMARKS

A unified and direct method of obtaining necessary conditions is thus obtained for optimal control problems. An important aspect which emerges from these considerations is that the minimum principles for differential, partial differential, integral, integrodifferential, etc. systems branch from a single basic conception.

REFERENCES

[1] L.S. Pontryagin et al., The Mathematical Theory of Optimal Processes, New York: Wiley, 1962.

[2] W.H. Ray and M.A. Soliman, "The optimal control of processes containing pure time delays-I Necessary conditions for an optimum," Chem. Engg. Sci., Vol. 25, pp 1911-1925, December 1970.

[3] S. Hiratsuka and A. Lchikawa, "Optimal control of systems with transportation lags," IEEE Trans. on Automatic Contr. Vol. AC-14, pp 237-247, June 1969.

[4] A.I. Egorov, "Optimal processes in distributed parameter systems and certain problems of invariance theory," SIAM J. Contr., Vol. 4, pp 601-661,1966.

[5] A.G. Butkovskii, Control of Distributed Systems, New York: Elsevier, 1969.

[6] R.R. Bate," The optimal control of Systems with transport lag," Advances in Control Systems, Vol.7, pp. 165-224, 1969.

[7] M. Athans," The status of optimal control theory and applications for deterministic systems," IEEE Trans. Automatic Contr., Vol. AC-11, pp 580-596, July 1966.

[8] R.C. Robinson, "A Survey of optimal control of distributed-parameter systems," Report No. ARL 69-0177, Aerospace Research Labs., November 1969.

[9] D.G.B. Edelen, Nonlocal Variationa and Local Invariance of Fields, New York: Elsevier, 1969.

[10] A.Ya. Dubovitskii and A.A. Milyutin,"Extremum problems under constraints," USSR Comput. Math. Phy., Vol. 5, pp 1-80, June 1965.

[11] A.M. Tsirlin,"Solving the optimal control problems by reduction to the simplest isopermetric problem," Eng; Cybernetics, Vol. 5, pp 134-143, May 1968.

[12] I.M. Gelfand and G.E. Shilov, Generalized Functions Vol. I, New York: Academic Press, 1964.

A STUDY OF DEMAND SCHEDULING IN A PARALLEL PROCESSING SYSTEM

Y. E. CHEN
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

V. Y. SHEN
Dept. of Computer Sciences
Purdue University
Lafayette, Indiana 47907

ABSTRACT

The processor assignment problem for a multiprocessing system with two classes of processors is investigated. A general scheduling strategy, demand scheduling, is defined as the strategy that allows no processor to become idle unless there is no process waiting for that kind of service. This paper demonstrates that based on this strategy, both the upper and lower bounds on system efficiency can be obtained systematically without the complete knowledge of the structure of job requests when certain criterions are satisfied. This information can be used to compute the scheduling characteristic which is a good measure of the fluctuation of the system efficiencies caused by different scheduling strategies.

I. INTRODUCTION

The problem of designing an optimal scheduling algorithm for multiprogramming and multiprocessing systems has proved to be very difficult. Several interesting results have been obtained dealing with the total length of execution [1,2], the execution anomalies [3,4], the swapping strategies [5], and others. In all the above papers the authors assumed that there was only one type of "active" system resources--the central processors. This paper describes a study on the bounds of system efficiencies of different scheduling algorithms in an environment with two classes of processors--the central processors and the peripheral processors.

For a multiprocessing system with two processors, a job can be modeled by an acyclic graph known as the resource request graph. A typical resource request graph is shown in Figure 1. The graph completely specifies the demands on the central processors (CP's) and the peripheral processors(PP's) and their precedence relations.

Let the multiprocessing system under study have m central processors and n peripheral processors. The maximum degree of system utilization in a time period T is defined as

$$U_{max} = mT + \alpha nT, \qquad (1)$$

where α is the cost ratio between the two classes of processors. The system efficiency E is defined as the ratio of the actual degree of system utilization vs. U_{max}.

The scheduler in a multiprocessing system has the responsibility of assigning processors to process the resource requests of the jobs. A general strategy used by the schedulers is one that allows a processor to become idle only if there is no request for this type of service. This strategy will be called the principle of demand scheduling. Although this is a very natural way of assigning processors, the following example shows that some deliberate delay in the processor assignment may result in higher efficiency for the system.

Example 1: Given a system with parameters m = n = 2, $\alpha = 1$, and T = 20. For the set of jobs as shown in Figure 2(a), the optimal schedule is shown by the Gantt chart in Figure 2(b).

The efficiency for the schedule in Figure 2(b) is 0.625. If we schedule the three initial CP requests as shown in Figure 2(c) and delay the assignment of PP2 for 1 time unit, we may obtain an efficiency as high as 0.813.

The use of the special delay of certain processors to improve the system efficiency as shown in Example 1 normally requires extensive trial-and-error approaches. Since the cost of this extra processing by the scheduler may be prohibitive, we shall only consider the effects of demand scheduling.

The system efficiency is a direct measure of the performances of different scheduling schemes. Since an optimum scheduling scheme for a certain set of jobs is not likely to be optimum for a different set of jobs, exhaustive enumeration seems to be the only way of finding the optimum scheme in a particular case. A partial solution to this problem is the definition of the scheduling characteristic as follows.

$$C = 1 - (E_L/E_U) \tag{2}$$

In equation (2), E_L and E_U are the lower and upper bounds of the system efficiency for a set of jobs under the principle of demand scheduling. It is evident that if C is small for a certain set of jobs, all scheduling schemes based on the principle of demand scheduling should have similar performances. This would eliminate the need of searching for the optimum scheme. The following discussions present two systematic methods to find the upper and lower bounds on the system efficiency for a given set of jobs.

II. THE DERIVATION OF E_U

It is necessary to determine first the resource requests of the jobs in the time period. This can be done by tracing the structure of each resource request graph and removing the parts which can only be processed after T time units. An example to show this procedure is shown in Figure 3. We shall not elaborate on the construction of such a simple algorithm and all resource request graphs are considered "trimmed" to fit in T.

A number of processors may be idled near the beginning of the time period if there are not enough initial requests or if the requests are mostly made to only one type of the processors. This initial idle time can be considered in finding an upper bound for the system efficiency. Since the system is essentially symmetric with respect to the two classes of processors, the following discussion is limited to the central processors.

When the CP's become idle, they are waiting for some non-initial CP requests to become available. The available time for each non-initial CP request is dependent on the actual schedule. However, we can find the "earliest available time" for each non-initial CP request by assuming that there is no delay in the execution of each request preceding the non-initial request being considered. For example, for the resource request graph shown in Figure 4, nodes b, c, and d are the non-initial requests. (Node a is available with the initial node hence is part of the initial request.) By tracing the graph, the earliest available time for nodes b, c, d are 9, 12, and 14 time units respectively.

We shall select m smallest "earliest available times" of the non-initial CP requests. If the total number of the non-initial CP requests is less than m, additional starting times of value T will be used to make up the difference. Conceptually, the m smallest "earliest available times" form the boundary of a region at the beginning of the CP schedule so that only the initial CP requests can be executed within the region. Therefore, the lower bound of the initial CP idle, I_{CI}, is the sum of the m earliest available times subtracting the sum of all the initial CP requests. If the result is negative, we shall consider that there is no initial idle. A similar procedure can be applied to derive I_{PI}, which is the initial PP idle time. If the number of the initial CP requests is less than or equal to m, there is only one way of executing the initial requests. This is true because all the requests must be initiated at time zero based on the principle of demand scheduling. A better lower bound of I_{CI} can be derived by taking the sum of any positive results of the pairwise subtraction so arranged that the largest of the m earliest available times match the largest initial CP request, and so on. We can derive I_{PI} using a similar method if the number of the initial PP requests is less than or equal to n.

Example 2: Given a system with parameters $m = 5$, $n = 7$, $\alpha = 1$ and $T = 10$. For the set of jobs in Figure 5, the information of the earliest available times of non-initial requests are placed at the right-hand side of the resource request graphs. The entries marked with asterisks are used to compute I_{CI}, and the entries marked with crosses are used to compute I_{PI}. Since the number of initial requests is less than the number of processors, we shall use pairwise subtractions for the initial idle times. The function POS(x) is defined to be x only if $x > 0$; otherwise it is zero.

$$I_{CI} = \text{POS}\ (4-9) + \text{POS}\ (4-4) + \text{POS}\ (4-3) + \text{POS}\ (3-2) + \text{POS}\ (2-0) = 4$$

$$I_{PI} = \text{POS}\ (5-7) + \text{POS}\ (5-5) + \text{POS}\ (4-4) + \text{POS}\ (4-3) + \text{POS}\ (4-2) + \text{POS}\ (3-0) + \text{POS}\ (2-0) = 8$$

Let the number of jobs in the set be k, each job has t_c CP request and t_p PP request in the time period T. Two upper bounds on the CP and PP utilizations can be formulated as the following:

$$U_{cmax} = \text{MIN}\ (mT-I_{CI},\ \sum_{i=1}^{k} t_{c_i}) \tag{3}$$

and

$$U_{pmax} = \text{MIN}\ (nT - I_{PI},\ \sum_{i=1}^{k} t_{p_i}) \tag{4}$$

where the function MIN(x,y) has the value of the lesser of the two arguments. An upper bound on the system efficiency is therefore

$$E_U = \frac{U_{cmax} + \alpha U_{pmax}}{U_{max}} \tag{5}$$

For the above example, E_U is 0.9.

III. THE DERIVATION OF E_L

A lower bound on the system efficiency can be derived by finding, based on the principle of demand scheduling, the minimum degrees of the CP and PP utilization, [6,7]. The combined result is then compared with the maximum degree of system utilization. Since the two bounds on the processor utilization were derived rather independently, it can be shown that the lower bound on the system efficiency so obtained is not reachable under most circumstances. We shall describe in this paper a systematic method to find a better bound when the set of jobs satisfy some criterions which are easy to identify.

A good schedule has plenty of overlapping in the processing of the CP and PP requests. In order to obtain a lower bound on the system efficiency, we shall construct several imaginary schedules that show the minimum overlapping of the two types of processes. As a first step, we shall construct an imaginary schedule S_I for the initial requests of the set of jobs only. The imaginary schedule S_I is to have the property that the system efficiency in any sub-interval starting at time zero is less than or equal to the system efficiency in the corresponding interval of the worst schedule S_W. From S_I we shall determine a time instant α and construct another imaginary schedule that includes the non-initial requests in a non-overlapping fashion after τ.

The two classes of initial requests can be scheduled independently. We shall consider the schedule for the central processors first. Assume we are given a system with three processors, and the initial requests are of lengths 1, 2, 2, 4, 5, 5, 7 time units. A schedule is shown in Figure 6(a). It is possible to prove, by an exhaustive method, that the schedule has CP efficiency lower than or equal to the efficiency of any other schedule under demand scheduling in any time interval between 0 and τ, $\tau \leq 8$. It is also possible to form a schedule as shown in Figure 6(b) and show that the schedule has efficiency lower than or equal to the efficiency of any other schedule in any time interval between 0 and some time instant $8 \leq \tau \leq 13$. Therefore it is in general not possible to construct a single schedule of the initial requests so that its efficiency is lower than or equal to any other schedule for any time interval starting at time equals zero. We shall therefore try to construct an imaginary schedule of the initial requests that has such a property.

For the initial requests, the principle of demand scheduling requires that no processor become idle until all the initial requests have been started. It is therefore possible to define a time instant t_e, which is the time that the first processor becomes idle. All the other initial requests have been either completed or already started at t_e. We shall refer to t_e as the "earliest termination time" of the initial requests, which is 5 for our example in Figure 6(a).

If the number of the initial CP requests is less than 2m, t_e can be determined by scheduling the m smallest initial requests at time equals 0, and schedule the rest at the earliest possible opportunities. Thus t_e is found to equal the length of the mth largest initial request. t_e is of course zero if the number of the initial requests is less than m. As the number of initial requests gets large, it becomes increasingly difficult to determine the exact value of t_e. However, a bound for t_e can be obtained by assuming that the m-1 largest initial requests are started at the same time instant the first processor becomes idle. Thus the bound is the sum of the initial requests, after excluding the m-1 largest ones, divided by m. In the example of Figure 6, it should be 4 2/3. Since all requests being scheduled have integral valued lengths, a better bound is the smallest integer greater than or equal to the above quotient. This makes the bound to be 5, which turns out to be the actual value of t_e for this example.

It is also of interest to find the "latest termination time" of the initial requests, which is referred to as t_ℓ. Based on the principle of demand scheduling, the difference between the time the first processor becomes idle and the time the last processor terminates should be less than or equal to the length of the largest request. An example is shown in Figure 6(b). The exact value of t_ℓ is again difficult to find unless the number of initial requests is small. A bound for t_ℓ can be obtained by assuming that the largest initial request is started at the same time instant all the other processors become idle. Thus the bound is the sum of the initial requests, after excluding the largest request, divided by m and then added by the length of the largest request. In the example of Figure 6, it should be 13 1/3. Since processes only terminate at integral time units, a better bound is the largest integer less than or equal to the above result.

This makes the bound to be 13, which turns out to be the actual value of t_ℓ for this example.

The imaginary schedule S_I is constructed by disregarding the actual lengths of the individual initial requests and forming as many processes that terminate at t_ℓ as possible, with no process terminating before t_e. The imaginary schedule for the example is shown in Figure 6(c). The general form of the CP schedule of S_I is shown in Figure 7(a). An actual schedule of the initial CP requests has the general form shown in Figure 7(b). It is evident that S_I has the same CP efficiency as any other schedule in any time interval between 0 and t_e. In the interval between t_e and t_ℓ, S_I has delayed certain processes as much as possible. Therefore S_I has CP efficiency lower than or equal to the efficiency of any other schedule of the initial requests in any time interval starting at time equals 0. Since the worst schedule S_W schedules both the initial and the non-initial requests, its efficiency is even higher.

As shown in Figure 7(a), it is possible to have an imaginary process terminating at t_m, $t_e < t_m < t_\ell$. Let the sum of the initial CP requests be S. The number of imaginary processes that terminate at t_ℓ is

$$n_\ell = \left\lfloor \frac{S - mt_e}{t_\ell - t_e} \right\rfloor \tag{6}$$

where [x] represents the largest integer less than or equal to x. If the quotient in Equation (6) is not an integer, then

$$t_m = S - mt_e - n_\ell\,(t_\ell - t_e). \tag{7}$$

An example is shown in Figure 6(c).

The imaginary PP schedule for S_I is constructed in a similar fashion. Given the earliest, middle, and latest termination times for the imaginary PP processes, we shall combine them with the termination times of the CP processes, sort them in ascending order, and rename them t_1 through t_6. (There may be as few as four values when the quotient of Equation (6) is an integer, or there may be more than six values when the number of the initial requests is less than the number of their corresponding processors for large systems.) It follows that the system efficiency of S_I in any interval starting at time equals zero is lower than or equal to the system efficiency of S_W in the same interval.

The next step in deriving a lower bound on the system efficiency is to consider the effect of the non-initial requests. They shall be introduced to S_I at a time instant τ, the value of which will be described later. The assumption is that there is no overlapping between the CP and PP processes after the time instant τ. Two imaginary schedules, S_A and S_B, may be constructed as shown in Figure 8, and we shall prove that the efficiency of one of them is a lower bound under certain circumstances.

Consider once more the worst schedule S_W. We shall divide it into subintervals such that there is no change of state (active and idle) for any processor within a subinterval. We can construct an imaginary schedule S_W' from S_W so that there is no overlapping between the CP and PP processes after a time instant τ. This is accomplished by distributing the CP work equally among the CP's and the PP work equally among the PP's for each subinterval after τ, and create two new subintervals with no overlapping. This conversion is shown in Figure 9.

The efficiency of the imaginary schedule S_W' is lower than or equal to S_W if each subinterval being converted expands or remains the same length ($\Delta t' \geq \Delta t$ in Figure 9). The following lemma describes such a condition.

Lemma 1. In an interval with m CP's and n PP's m' CP processes and n' PP processes, the modification technique as shown in Figure 10 causes the interval to expand or remain the same length if and only if

$$\frac{m'}{m} + \frac{n'}{n} \geq 1 \tag{8}$$

Proof: Let the length of the original interval be Δt and the length of the modified interval be $\Delta t'$. Then

$$\Delta t' = \frac{m'}{m} \Delta t + \frac{n'}{n} \Delta t$$

Relation (8) is the necessary and sufficient condition for $\Delta t' \geq \Delta t$.

It is possible to relate the efficiency of S_W' to one of the efficiencies of S_A and S_B if all subintervals in S_W satisfy Relation (8). We shall first group the subintervals of S_W' so that the CP processes are next to each other, and the PP processes are next to each other. If $m > n\alpha$, to schedule all the PP processes ahead of the CP processes produces lower efficiency in the interval between τ and T, since the PP's cost less. Thus the efficiency of the imaginary schedule S_A is a lower bound, for it has lower efficiency in both the intervals from 0 to τ and from τ to T. On the other hand, if $m < n\alpha$, the efficiency of the imaginary schedule S_B is a lower bound. When $m=n\alpha$ both the efficiencies of S_A and S_B are bounds and the higher one should be used.

The determination of τ is not crucial. It should be one of t_1, t_2, t_3, t_4, t_5, and t_6. The one that gives the highest lower bound can be chosen as follows. The six time instants defined earlier form six intervals at the beginning of the time period. (There may be more or less intervals than six as described earlier.) The time instant τ is t_i if the interval from t_i to t_{i+1} has higher efficiency when it is fully scheduled with one type of processes. Let the number of CP processes in an interval be m', and the number of PP processes in an interval be n'. In S_A we select τ if the interval immediately to the right no longer satisfies $m' + n'\alpha > n\alpha$. Similarly, the condition in S_B is $m' + n'\alpha > m$.

To summarize the procedure, we first construct the imaginary schedule S_I for the initial requests of the set of jobs. One of the time instants t_1 through t_6 is chosen to be τ. Based on the comparison of m and $n\alpha$, one of the imaginary schedules S_A and S_B is constructed using all the remaining resource requests. The efficiency of the imaginary schedule is a lower bound on the system efficiency if Relation (8) is satisfied by all the subintervals of S_W. Although the condition seems abstract, it in fact corresponds to a moderate or heavily loaded system. The following are some criterions which can be checked against the set of jobs easily. If one of them is satisfied, it implies that Relation (8) is satisfied by S_W.

Criterion 1. In the resource request graphs of the jobs, there are at least MAX(m,n) jobs with the lengths of the longest paths equal to T.
Criterion 2. The time instant τ as determined is greater than or equal to T. This indicated that there are sufficient initial requests.
Criterion 3. Let the jobs be ordered descendingly according to the lengths of their longest paths in the resource request graphs. Remove the jobs one by one, and let m'' and n'' be the number of the CP and PP starting jobs respectively. The removing procedure ends before $m''/m + n''/n \geq 1$. The sum of the lengths of the longest paths belonging to the remaining jobs is greater than or equal to mT + nT.

Example 3: For the system as described in Example 2, and the set of jobs shown in Figure 5. Criterion 1 is satisfied by the presence of 8 jobs with lengths of the longest paths equal to 10. Since the numbers of initial

CP and PP requests are less than their corresponding processors, there is only one way to execute the initial requests. The schedule (S_I) is shown in Figure 10(a). τ is determined to be 3 using the conditions described. Since $m < n\alpha$, and the total sum of CP requests is larger than or equal to $m\tau$, the remaining interval is covered by CP processes. The imaginary schedule S_B is shown in Figure 10(b). Therefore,

$$E_L = \frac{46 + 14 \times 1}{50 + 70 \times 1} = 0.5$$

The scheduling characteristic is

$$C = 1 - \frac{0.5}{0.9} = 0.44$$

IV. CONCLUSION

We have described in this paper two methods to determine the upper and lower bounds of the system efficiency for a set of jobs under demand scheduling. The method to determine an upper bound on the system efficiency in a given time period uses information on the total CP and PP requests of the jobs in the time period, the lengths and types of the initial requests, and the precedence relations at the beginning of each job's resource request graph. When the set of jobs satisfies certain criterions, the method to determine a lower bound used the total CP and PP requests and the lengths and types of the initial requests. The two bounds may be used to calculate the scheduling characteristic, which is a measure of the relative performances of all scheduling schemes using the principle of demand scheduling. The bounds so derived are reachable. Since the length of the time period under consideration is arbitrary, it is possible to obtain reasonable estimates on the system efficiency either dynamically or statically, if the structures of the jobs are known.

This study, together with the previous work in References 6 and 7, give a general behavior of the scheduling characteristic under demand scheduling. The scheduling characteristic is small for a lightly loaded system since all requests may be processed before the end of the time period. It increases with the load as E_U increases faster than E_L until it reaches 1. The highest value for C is reached in a moderately loaded system in which the sum of all the requests is approximately equal to the maximum system utilization. This is the time that a good scheduling strategy should be employed. The scheduling characteristic then decreases as E_L increases with the load, until it also reaches 1. This corresponds to an overloaded system for which just the initial requests of the jobs are enough to saturate the processors.

V. REFERENCES

1. Hu, T. C., Parrallel Sequencing and Assembly Line Problems. Operations Res. 9 (November 1961), 841-848.
2. Muntz, R. R., and Coffman, E. G., Optimal Preemptive Scheduling on Two-processor Systems. IEEE Trans. Computers C-18 (November 1969, 1014-1020.
3. Graham, R. L., Bounds on Multiprocessing Timing Anomalies. SIAM J. Appl. Math. 17 (March 1969), 416-429.
4. Manacher, G. K., Production and Stabilization of Real-time Task Schedules. J. ACM 14 (July 1967), 439-465.
5. Denning, P. J., The Working Set Model for Program Behavior. C. ACM 11 (May 1968), 323-333.
6. Chen Y. E., and Shen, V. Y., Scheduling Problem in a Multiprocessing Environment with Two Classes of Processors. Proc. 5th Princeton Conf. on Information Sciences and Systems (March 1971), 110-115.

7. Chen, Y. E., and Shen, V. Y., On the Bounds of System Efficiency in a Multiprocessing Environment with Two Classes of Processors. Tech. Rep. TR-EE 71-7 (April 1971), School of Elec. Engineering, Purdue University, Lafayette, Indiana.

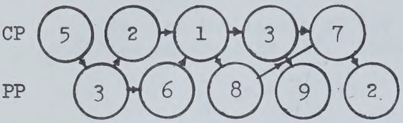
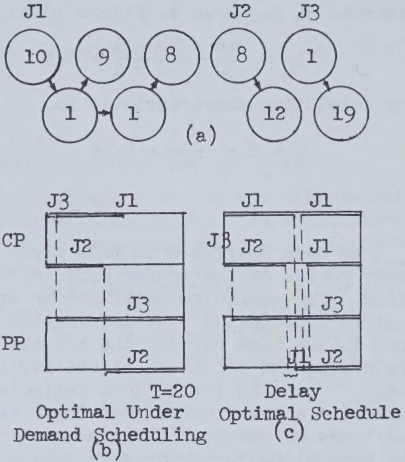

CP
PP
5
2
1
3
7
3
6
8
9
2

Figure 1


J1
J2
J3
10
9
8
8
1
1
1
1
12
19
(a)
J3
J1
CP
J2
J3
PP
J2
T=20
J1
J1
J3
J2
J1
J3
J1
J2

Optimal Under Demand Scheduling (b)

Delay Optimal Schedule (c)

Figure 2

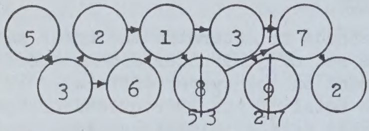

5
2
1
3
7
3
6
8
9
2
5 3
2 7

Figure 3

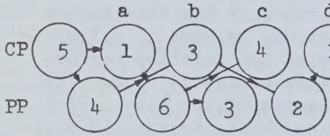

a
b
c
d
CP
5
1
3
4
1
PP
4
6
3
2

Figure 4

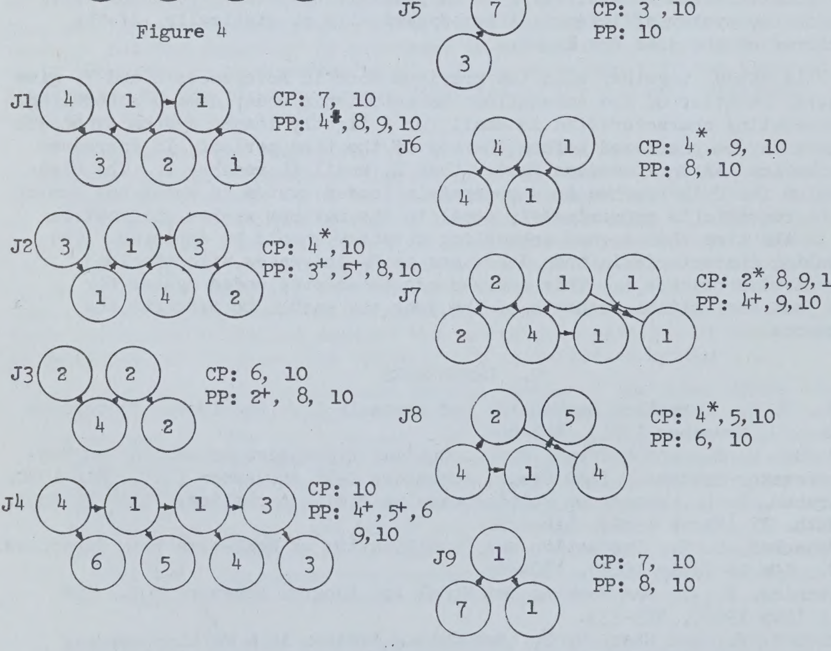
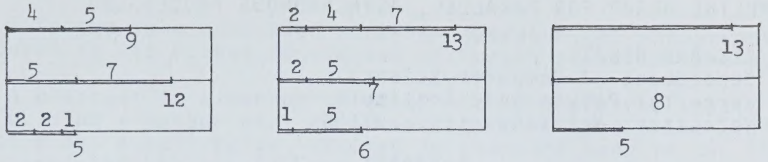
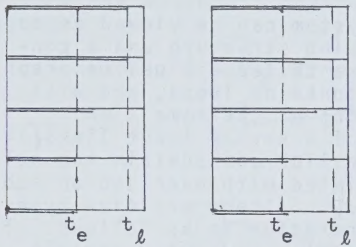

J1
4
1
1
3
2
1
CP: 7, 10
PP: 4*, 8, 9, 10
J2
3
1
3
1
4
2
CP: 4*, 10
PP: 3+, 5+, 8, 10
J3
2
2
4
2
CP: 6, 10
PP: 2+, 8, 10
J4
4
1
1
3
6
5
4
3
CP: 10
PP: 4+, 5+, 6
9, 10
J5
7
3
CP: 3*, 10
PP: 10
J6
4
1
4
1
CP: 4*, 9, 10
PP: 8, 10
J7
2
1
1
2
4
1
1
CP: 2*, 8, 9, 10
PP: 4+, 9, 10
J8
2
5
4
1
4
CP: 4*, 5, 10
PP: 6, 10
J9
1
7
1
CP: 7, 10
PP: 8, 10

Figure 5


4
5
9
5
7
12
2 2 1
5
2 4
7
13
2 5
7
1 5
6
13
8
5

(a) The earliest termination time (b) The latest termination time (c) The imaginary schedule

Figure 6


t_e
t_ℓ
t_e
t_ℓ

(a) The imaginary schedule (b) An actual schedule

Figure 7

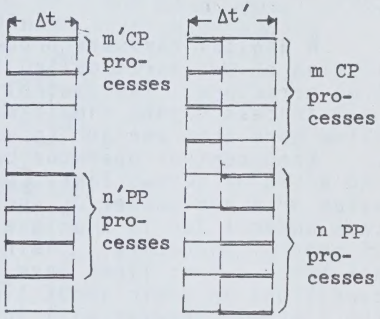

Δt
m′CP processes
n′PP processes
Δt′
m CP processes
n PP processes

Figure 9

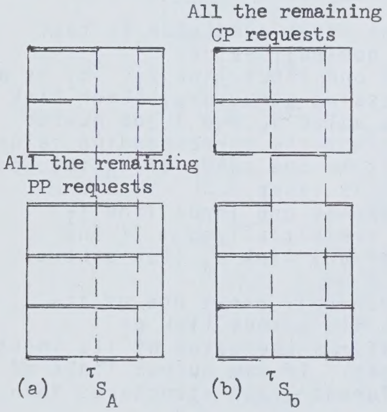

All the remaining CP requests
All the remaining PP requests
(a)
τ
S_A
(b)
τ
S_b

Figure 8

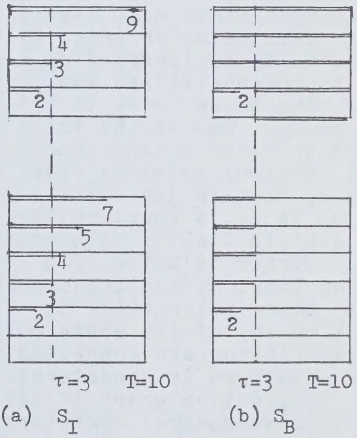

9
4
3
2
7
5
4
3
2
2
2
τ=3
T=10
τ=3
T=10
(a) S_I
(b) S_B

Figure 10

A PIPELINE GRAPH FOR PARALLEL, ASYNCHRONOUS PROCESSING

RICHARD RIBAK
Department of Computer Science
Worcester Polytechnic Institute
Worcester, Massachusetts 01609

ABSTRACT

A parallel asynchronous control structure is developed which is called a pipeline graph. It controls the simultaneous processing of N independent jobs. The pipeline graph is developed to be used in the modular design of complex computer systems.

SUMMARY

A digital hareware processing system can be viewed as consisting of two parts called a processing structure and a control structure. The control structure called a pipeline graph will process N jobs simultaneously, contains loops, and will allow more than one job to enter a loop at the same time.

Each control operator consists of a set of input links $\{i_j\}$ and a set of output links $\{p_j\}$. Each link can contain the null value or a job number J_k where associated with each job or subtask under a job is a unique number J_k . There are five types of control operators. Their basic operation is as follows. If all their output links have the null value and under certain conditions on their input links (to be specified subsequently), the control operator will initiate the processing structure. When processing is completed it will set its input links to the null value and its output links to the job number J_k previously in its input links.

(a) Function Operator (Fig. 1)--The input condition is that all input links must have the same non-null value.

(b) Data Selector (Fig. 2)--It has one input link i_1. b_1 is a data value monitored from the processing structure. Each link a_k is preinitialized and contains a value V_k and a job number j_k. The value in i_1 is matched against the corresponding value j_b in a_b. Then if the value in b_1 is in the range of V_b the yes path is taken otherwise the no path is taken.

(c) Control Selector (Fig. 3)--There is one input link i_1. The a_k contain job numbers and are preinitialized. If the value in i_1 is equal to the value of any link a_j then control is sent to link p_1 otherwise to link p_2.

(d) Merger Selector (Fig. 4)--It randomly picks one of its input links i_k and transforms it to its output link p_1.

(e) Distributor (Fig. 5)--It transforms the value of its input link to any of its empty output links. If the output links of a distributor are connected to different graph structures then the transform is nondeterministic.

A pipeline graph is defined as the interconnection of the above five control operators.

A control operator in standard form has a set of input queues instead of input links. When the identical number has appeared on each of its input queues, that number will be transformed to the set of output links of the operator.

Theorem 1: A pipeline graph whose operators are in standard form is output functional.

Theorem 2: A loop free pipeline graph in standard form will not hangup.

A hangup is a condition where an operator can not transform its input to its output regardless of future states of the system.

A p-merger is a merger selector with a preferred input link b. The p-merger will not transform the input links $\{i_j\}$ unless b has a null value. A loop in standard form is one that is entered through the $\{i_j\}$ of a p-merger and the loop itself enters through preferred link b and leaves through the output.

Theorem 3: Loops in standard form will not hangup.

A hierachial structure is a graph which controls a second graph. A major application of a hierachial structure is the design of a control processor in terms of a set of processing instructions which are common to most high level programming languages.

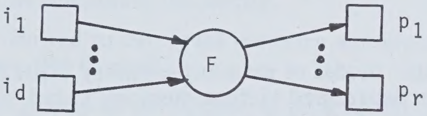


Fig. 1 Function Operator

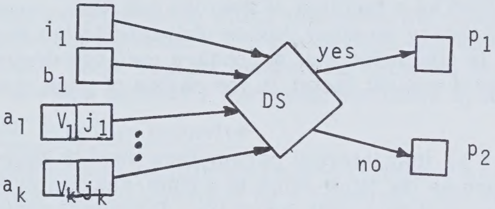


Fig. 2 Data Selector

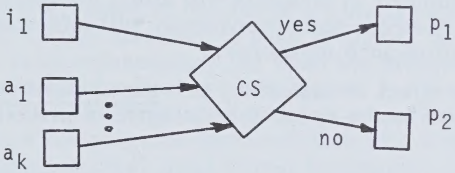


Fig. 3 Control Selector

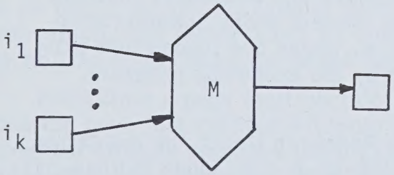


Fig. 4 Merger Selector

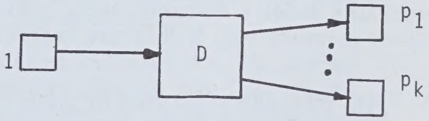


Fig. 5 Distributor

THE WASTED SPACE-TIME INTEGRAL AND PAGE SIZE IN THE DESIGN OF VIRTUAL MEMORY COMPUTER SYSTEMS*

by

Erol Gelenbe
and
Paolo Tiberio †

Department of Electrical and Computer Engineering and
Computer, Information and Control Engineering Program
The University of Michigan, Ann Arbor, Michigan 48104

Abstract. A stochastic model of primary memory utilization for compute-bound jobs in a multiprogrammed virtual memory paged computer system is developed. The effect of page size on the wasted portion of the space-time integral of primary memory is studied, and the page size minimizing this waste is determined as a function of average job size, mean time beteen page faults, secondary memory device characteristics and other relevant parameters. It is concluded that secondary memory device characteristics should be the dominant factor in the choice of page size.

1. Introduction

It is well known [1] that internal parameters used in scheduling a computer system (such as the time-slice in a time-sharing system) will affect the service received by a user program. Similarly design parameters of virtual memory [2] systems will affect the efficient utilization of installed storage devices, and consequently will influence various system oriented performance measures.

In this paper the effect of page size, in a paged virtual memory machine [2], on waste in the space-time integral of primary memory (PM) is investigated.

In typical memory systems [2] a program is partitioned into equal sized pages of units of space (e.g. words or bytes). Main memory is divided into equal sized page frames. Some of the pages of a given program will reside in some of the primary memory page frames, while the remaining pages will be in some secondary storage device, though it is conceivable that some pages may be in primary and secondary memory simultaneously. We shall assume that the system under consideration operates under a demand paging policy, i.e. pages are loaded into primary memory only if they are referenced by the executing program.

---

* This work was supported by a Rackham Research Grant, by the Office of Naval Research under Contract N00014-67-A-0181-0036 (NR049-311) and by a NATO Science Fellowship to P. Tiberio.

† On leave from the Centro Studi per Interazione Operatore-Calcolatore Instituto di Elettronica, University of Bologna, Italy. This author's contribution is dedicated to Professor U. Tiberio.

In the system we are considering here, PM is used by the CPU to fetch and store instructions and data. It is also by far the most expensive memory unit in cost/bit. Therefore it constitutes the memory space which must be utilized most efficiently. The waste of its space-time integral (WSTI) is a measure of how efficiently it is being utilized. The only attempts to determine page size analytically [2] , [3] ,[4] have stressed PM space wasted due to internal and table fragmentation [2] . These approaches have overlooked the importance of the secondary memory (SM) in PM utilization and hence on the choice of page size. In this paper the effect of page replacement policies [2] on the WSTI is not considered directly. Rather, the mean time between consecutive page faults [2] is introduced as a parameter, so that the effect of page replacement policies may be evaluated indirectly.

This study is limited to two level memory systems, i.e., those consisting of PM and one SM device.

In Section 2 we give a broad outline of program system behavior as it relates to our study. Section 3 contains the mathematical model upon which this study is based. In Section 4 we present some numerical examples giving the page sizes which minimize the wasted space-time integral for different secondary storage devices and assumed distributions of program size. These results indicate an optimum page size close to the page size in some presently operating computer systems.

2. System and Program Behavior

In the framework of multiprogrammed computer systems the execution of a program is characterized by a repetive cycle consisting of computation, input-output, and time spent in the CPU queue, in that order. These cycles are interrupted at the end of a time slice (a predetermined maximum total CPU time allocated to each program, at the end of which the program has to be rescheduled), or when the program encounters a symbolic STOP statement. During computation, the program is receiving attention of the CPU. The time spent in input-output may consist of a period during which the program communicates with a time-sharing terminal, with files, or other input-output equipment, or the time spent in page-wait from secondary storage devices [2] such as drum, large capacity storage (LCS) or extended core storage (ECS). Our concern here is with this last item only. Page-wait will generally consist of time spent in a queue waiting for the secondary storage device to become available, followed by the time necessary to transfer a page. The magnitude of its effect on overall system performance will depend on the frequency with which page faults occur as opposed to other input-output operations, as well as on the duration of the page-wait.

Each time a page fault occurs, it will be assumed that the system will load the page just referenced from secondary storage into PM. If the program already occupies the maximum number of pages allotted to it by the system, the incoming page will occupy a page frame previously occupied by another page of the same program. Otherwise it will occupy a page frame without forcing a page of the same program out of primary memory. The time spent in page-wait multiplied by the number of pages occupied by the program during page-wait will constitute one part of the wasted space-time integral. The total wasted space-time integral will include the effects of internal and table fragmentation [3] , [2] (or wasted space due to the "rounding off" effect in assigning an integer number of pages to a program, and the space wasted in page tables indicating the whereabouts of each program page, respectively), as well as the influence of the time between fauls.

3. The Model

Several studies have indicated the program behavior sketched in Figure 1 [6] ,[7] . The program, starting initially with no pages in PM, generates page faults at a near constant rate (and these are loaded into PM under a demand paging policy) until a number W of pages are in PM. The page fault rate diminishes considerably thereon. If the page size is T, W corresponds to

$$K = WT \tag{1}$$

amount of memory space. To simplify the terminology, we shall use a word as the basic memory unit. We assume that the program operates with K words of its address space in PM. From Figure 1 we deduce that the time between page faults m, up to the time when the program has K words in PM, is given by

$$m = \frac{\alpha_0 T}{K_0} \tag{2}$$

After K words have been loaded into PM, the mean time between successive page faults will be $M \geq m$, a constant. Figure 1 reflects the behavior of the program which is able to acquire the pages it requests. When it is constrained to swap some pages for one it has just referenced, the mean time between faults cannot be deduced from Figure 1. Other studies, principally those of Belady and Kuehner [8] , and Denning [9] , relate the page fault rate to the PM space available to the program.

When a page fault occurs, the page request is placed in the queue of the secondary memory (SM) as shown in Figure 2, to wait for service. It is assumed that page requests constitute a Poisson arrival process [10] . Although we are tracing the behavior of a single program, page requests may be generated by each program in the system so that these may be viewed collectively as a random process. Arguments for the validity of the Poisson assumption have been made on the basis of measurements on the University of Michigan Time-Sharing System [11] , [12] . This assumption simplifies the mathematical model. The service mechanism of the SM queueing system will depend on the SM device chosen. We will consider three devices in this study: drum, LCS, and ECS. We shall discuss the models for these devices in detail later. For the present, let us denote the time elapsed between a page request and when computation resumes on the program by t.

Let W_i be the number of pages in memory when the i-th page fault occurs. Thus $W_1 = 0$, and

$$W_{i+1} = \begin{cases} W_i + 1 & \text{if} \quad W_i < W \\ W & \text{if} \quad W_i \geq W \end{cases}$$

We assume that i = 1 when the first page of the program is requested. Since the program acquires an extra page frame as soon as a page fault occurs if $W_i < W$, the total WSTI during page faults up to the time that the W-th page is in PM is given by

$$S_{F1} = \sum_{i=0}^{W-1} (W_i + 1) \cdot T \cdot t = T \cdot t \cdot W \cdot (W+1)/2$$

or from (1)

$$S_{F1} = K \cdot t \cdot (W+1)/2 \qquad (3)$$

By this time, the program has a remaining computing time of $(C - \alpha) \geq 0$, during which page faults will occur each M time units. This will give a space-time integral waste of

$$S_{F2} = K \cdot t \cdot (C - \alpha)/M \qquad (4)$$

during page faults after W pages have been acquired.

The pagination of a program necessitates keeping a page table (PT) indicating the whereabouts of each program page, for each program. Let x be the program size. Then the PT size is

$$\left\lceil \frac{x}{T} \right\rceil \cdot b \qquad (5)$$

where b is a constant and $\lceil y \rceil$ is the least integer greater than or equal to y. In (5) the implicit assumption is made that the virtual memory is not segmented [2] as well as paged, because then each program segment would be paged, rather than the whole program address space of size x. The PT will be maintained in PM throughout the time a program is active. Disregarding the time before the program obtains its first page frame and after it has obtained a total computing time C, the total active period of the program is

$$\tau = C + W \cdot t + t \cdot (C - \alpha)/M \qquad (6)$$

so that table fragmentation contributes

$$S_T = b \cdot \left\lceil \frac{x}{T} \right\rceil \cdot (C + W \cdot t + (\frac{C - \alpha}{M}) \cdot t) \qquad (7)$$

to the total WSTI. Due to the assignment of an integral number of pages to the program, space is wasted in internal fragementation. This would normally be concentrated in one program page, and is given by

$$T \cdot \left\lceil \frac{x}{T} \right\rceil - x \qquad (8)$$

However, since the times when this page is contained in PM are unknown, it has been assumed that internal fragmentation is distributed among all program pages, so that the amount per page is

$$P = T - x / \left\lceil \frac{x}{T} \right\rceil \qquad (9)$$

With this assumption, internal fragmentation in PM is proportional to the number of page frames occupied by the program. Its contribution to total

WSTI is given by

$$S_I = \sum_{i=1}^{W-1} i \cdot P \frac{\alpha \cdot T}{K} + (C-\alpha)\cdot W \cdot P \tag{10}$$

During page-wait the total space occupied by a program is considered to be wasted. Therefore S_I is only calculated over the computing time. Using (1), (10) is rewritten as

$$S_I = \frac{\alpha \cdot P}{2} \cdot (W-1) + (C-\alpha)\cdot W \cdot P \tag{11}$$

The total WSTI for a program whose address space contains x words is then

$$S = S_I + S_{F1} + S_{F2} + S_T \tag{12}$$

Let us now examine the time t elapsed from the occurrence of a page fault to resumption of computation (attention of the CPU) to the program. Evidently, this time is a random variable. We shall henceforth use t to denote the average value of this time. t may be written as

$$t = t_{CPU} + t_{SM} + t_{SUP} \tag{13}$$

where t_{SUP} is the time the operating system will spend in supervisory activities as a result of the fault, t_{SM} is the time spent waiting for the page, and t_{CPU} is the time the program spends in the CPU queue waiting for the attention of the CPU after it has received a page.

We assume that t_{CPU} is independent of T. t_{SM} will depend on T, as well as the SM device. Therefore, we examine it for the drum, ECS, and LCS separately. t_{SUP} is also assumed independent of T.

The times elapsed between consecutive page faults are assumed to be independent, identically distributed random variables with exponential distribution and mean M so that the arrivals to the SM queue of Figure 2 constitute a Poisson process with parameter 1/M. Thus the ECS and LCS can be modelled as M/D/1 queues (Poisson arrival with constant service time) since they are random access devices. For the drum, the sectored drum model of Coffman [13] , [14] will be used.

Extended Core Storage (ECS) This is a magnetic core memory equipped with special data transmission facilities which allow it to deliver b words per PM cycle after an initial access time a [2] . Thus, the time to read or write a page is

$$a + \frac{T}{b} c \tag{14}$$

where c is the PM cycle time. Typically a = 3 μsec, c = 1 μsec, and b = 10 words. Thus, with the assumption that page requests constitute a Poisson arrival process with parameter λ, the ECS can be treated as an M/D/1 queue [10] , for which we immediately write .

$$t_{SM} = t_{ECS} = \frac{1}{M}(a+Tc/b) \left[1 + \frac{1}{M} \frac{a+cT/b}{2\left(1-\frac{1}{M}(a+Tc/b)\right)} \right] \tag{15}$$

With an empty queue, we obviously obtain

$$t_{ECS} = a + \frac{T}{b}c \tag{16}$$

Large Capacity Storage This is a core memory whose cycle time is considerably lower than that of PM. The time to read or write a page is

$$Tz \tag{17}$$

where z is typically 10 μsec. It is modelled as an M/D/1 queue so that

$$t_{SM} = t_{LCS} = \frac{1}{M}\,(Tz)\left[1 + \frac{Tz/M}{2(1-Tz/M)}\right] \tag{18}$$

For an empty queue

$$t_{LCS} = Tz \tag{19}$$

Drum The drum considered here is of the sectored type as used in paged systems [11] . Each page is addressed by a sector and track number, and the drum has to be rotated to the proper sector portion for a page read or write. The page requests are placed in a queue corresponding to their sector so that there is a ququе for each drum sector. This mechanism can be implemented in software or by some special purpose hardware. Let w be the number of words per track of the drum and let $\mathcal{J}$ be the time for a complete drum rotation. Then

$$\frac{T\mathcal{J}}{w} \tag{20}$$

is the time necessary to a page read or write after the drum has been positioned. It is assumed that there are N sectors in the drum and that page requests are equally likely to occur for any sector. The service time is obtained via an expression from [13] , [14] . Taking note of the fact that [14] contains an error as well as [13] and after having corrected it we obtain:

$$t_{SM} = t_{DRUM} = NMp\left(\frac{N+2}{2N}\right) + \frac{p^{2}NM}{\;} - \frac{\mathcal{J}}{N} \tag{21}$$

where

$$p = \frac{\mathcal{J}}{MN} \tag{22}$$

4. Computational Results

A FORTRAN IV program has been written which to compute S of (12) for different values of page size T, mean time between faults M,

t_{SUP}, t_{CPU}, and different secondary memory devices as well as program size distributions. Some of the results obtained have been plotted in Figures 3, 4, and 5, with typical values of the parameters and program size with Erlang probability density of mean 20,000 words. The variance of this distribution has a negligible effect of the curves. It is seen that the time spent in the CPU queue has an important influence on the page size minimizing the WSTI in all cases. Interestingly enough, the optimum page size for drum and ECS are nearly the same, while they are smallest for LCS. This is explained as follows. The latency effect forces pages to be large in a drum, since small pages do not affect the waste during page faults appreciably but do increase table sizes. In the ECS, page transfer time in comparable to t_{SUP} so that here too a small page size does not reduce the WSTI. The LCS is thus the only device in which the time to bring a page into memory is most directly influenced by the page size. In all cases t_{CPU}, the time spent in the CPU queue after a requested page has been obtained, should influence the choice of page size, and as seen in Figure 4 it is especially so with ECS.

In all cases, a smaller M implies a smaller optimum page size. This can be explained by the necessity to be able to give faster service at the SM device by reducing page size when the SM queue is longer.

Our main conclusions for compute-bound jobs are:

1) Page size should be determined primarily on the basis of the SM device and of the congestion at the CPU.

2) A page size of 512 or 1024 is optimum for systems with a paging drum, confirming current practice.

3) Smaller page sizes, of 128 or 256 words, are justified with LCS, but ECS requires larger pages for efficient operation.

4) As long as t_{CPU} is small, a faster SM device (e.g. ECS) improves PM utilization. This improvement is nullified when t_{CPU} becomes significant.

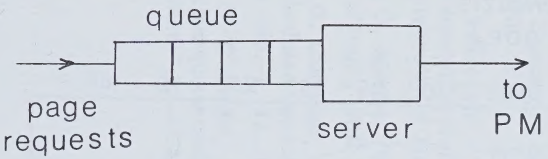

queue
page requests
server
to PM

FIGURE 2

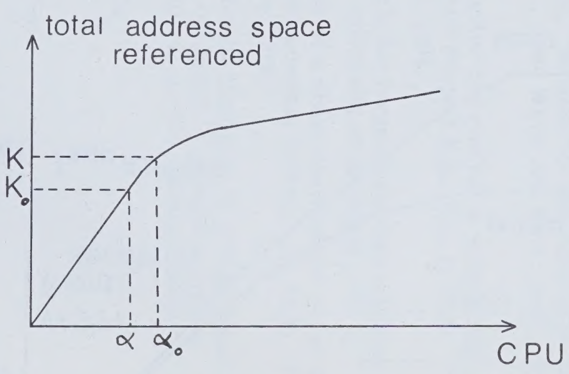


FIGURE 1

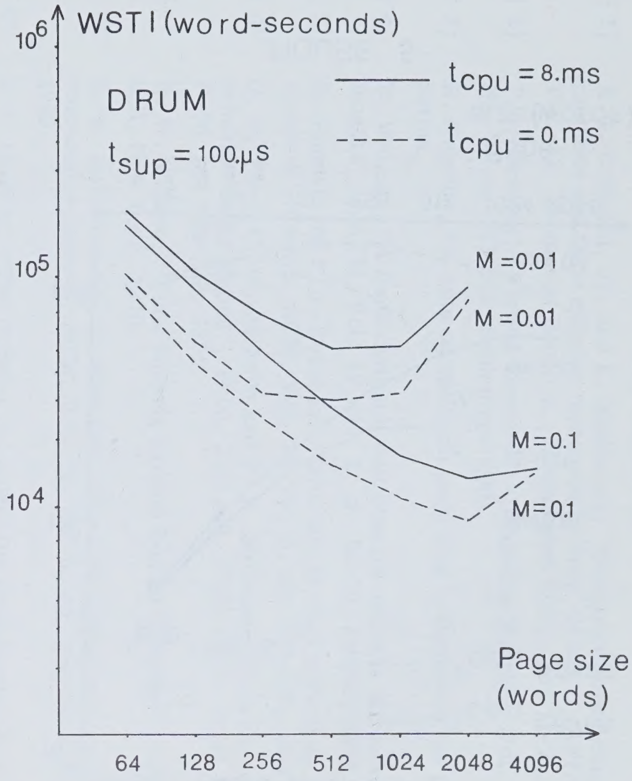


FIGURE 3

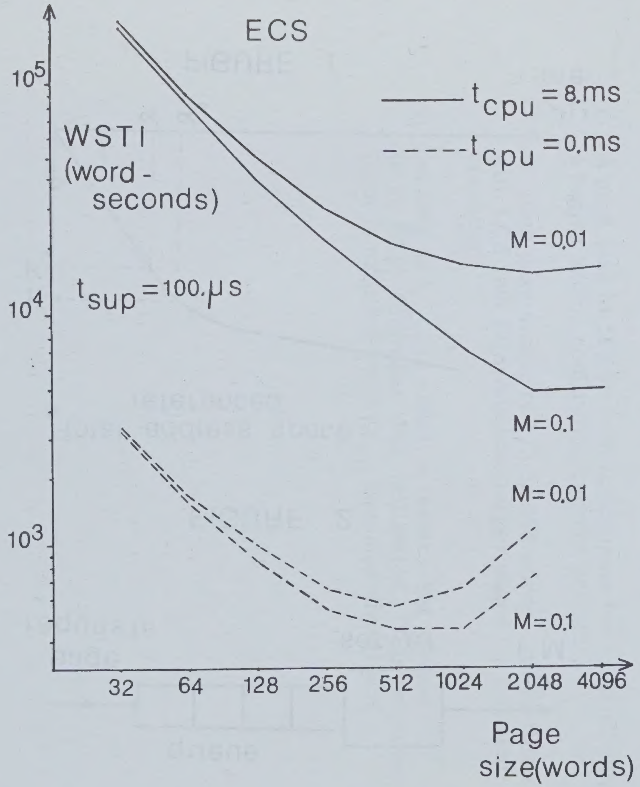


FIGURE 4

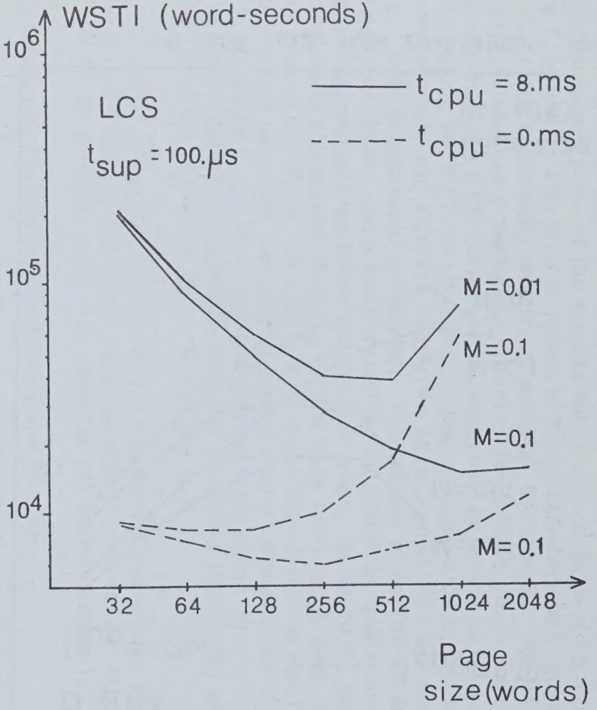


FIGURE 5

REFERENCES

[1] E.G. Coffman, Jr. and L. Kleinrock, "Computer scheduling methods and their counter measures", Proc. of SJCC, AFIPS, Vol. 32, Thompson Book Co., Washington D.C., pp. 11-21, 1968.

[2] P.J. Denning, "Virtual memory", Computing Surveys, Vol. 2, No. 3, pp. 153-183, September 1970.

[3] E. Gelenbe, "A note on optimum page size", submitted for publication.

[4] E. Wolman, "A fixed optimum cell size for records of various lengths", J.ACM, Vol. 12, No. 1, pp. 53-70, January 1965.

[5] B. Randell and C.J. Kuehner, "Dynamic storage allocation systems", Comm. ACM, Vol. 11, No. 5, pp. 297-305, May 1968.

[6] G. H. Fine. C. W. Jackson and P. V. McIsaac, "Dynamic program behavior under paging", Proc. 21st Nat. Conf. ACM, ACM Pub. P-66, pp. 223-228, 1966.

[7] R. W. Watson, "Time Sharing System Design Concepts", McGraw-Hill, New York, 1970.

[8] L. A. Belady and C. J. Kuehner, "Dynamic space-sharing in com computer systems", ACM, Vol. 12, No. 5, pp. 282-288, May 1969.

[9] P.J. Denning, "The working set model for program behavior", Comm. ACM, Vol. 11, No. 5, pp. 323-333, May 1968.

[10] A.B. Clarke and R.L. Disney, "Probability and Random Processes for Engineers and Scientists", John Wiley and Sons, New York, 1970.

[11] T. Pinkerton, "Program behavior and control in virtual storage computer systems", CONCOMP Project Rep. No. 4, The University of Michigan, April 1968.

[12] C. G. Moore, "Network models for large-scale time-sharing systems", PhD Thesis, The University of Michigan, 1971.

[13] E.G. Coffman, Jr., "Analysis of a drum input-output queue under scheduled operation in a paged computer system", J. ACM, Vol. 16, No. 1, 11. 73-90, January 1969.

[14] E.G. Coffman, Jr., Correction to "Analysis of a drum input-output queue under scheduled operation in a paged computer system", J.ACM, Vol. 16, No. 4, p. 646, October 1969.

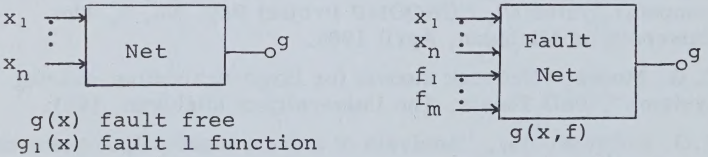
TEST GENERATION USING A NETWORK FAULT EQUATION*

JAMES W. GAULT
Electrical Engineering Department, North Carolina State
University, Raleigh, N. C. 27607

Abstract: The behavior of a digital network with a set of faults may be described, as suggested by Shertz [9], with a binary Boolean equation which is a function of both inputs and faults. Equations describing the network for single or multiple fault isolation or detection can be derived. The desired equation which results may then be directly analyzed to provide a set of appropriate tests.

INTRODUCTION

The derivation of tests for digital networks is dependent upon the particular structure of the network as well as the function realized. Much of the difficulty of deriving test inputs for a network revolves around the simultaneous modeling of these two properties. Many investigations have dealt with the detection and isolation of faults and have handled this interdependence in various ways [1][2][5][8]. Poage [5] was able to incorporate the total behavior into an equation description which was three valued. Shertz [9] then showed that similar information could be carried in a single binary function. This equation transforms the problem from an m+1 equation set of n variables, for m faults, to a single equation involving m+n variables as shown pictorially in figure 1.1.



$g(x)$ fault free
$g_1(x)$ fault 1 function
⋮
$g_m(x)$ fault m function

a. original network and equations

b. equivalent network and single equation

Figure 1.1. Transformed Network

This approach gives a binary, Boolean equation which need not be carried in any specific form in order to preserve the structural dependence of the fault behavior. The purpose of

---

*This work was partially sponsored by the National Science Foundation under Grant GK-5612.

this paper is to show how this equation can be used to directly obtain a test set for the purpose of isolation or detection of single or multiple faults.

Section two of this paper describes the nature of the equations formed and their derivation while section three discusses the test set selection.

2. GENERATION OF THE FAULT EQUATION

This section provides a brief description of the types of equations which are used and the way in which they are derived. The method of the next section is closely tied to the approach presented here.

Shertz [9] has indicated a procedure for deriving the fault equation which is denoted ZF here. The purpose now is to extend this derivation and then show how the desired test sets can be obtained directly from the equation representation.

A network, as for example figure 2.1a, can be viewed as a connected set of fault classes, figure 2.1b, as suggested by Shertz [9]. Here the original network is carried as a connected set of nodes and node descriptions. The node descriptions are given as disjunctive cubical expressions [3][4][6][8]. The faults for any particular node are obtained from the connected fault set description. The cubical node descriptions for the running example are given in figure 2.1c. Proceeding from the output to the input, nodes are injected [4][6] in turn until a fault function for the output Z is obtained, i.e. ZF(f,x) of figure 2.1d. Here f is a fault variable vector and x is an input variable vector as in figure 1.1b.

ZF is a one for any fault condition and input condition for which the original network would also give a one. It is convenient to separate those conditions for which the output is a one due to faulty behavior from the fault free behavior. This is done by defining

$$ZF1 = ZF \cap \overline{Z} \qquad (2.1)$$

$$ZF0 = \overline{ZF} \cap Z. \qquad (2.2)$$

It should be pointed out that

$$ZF = \overline{\overline{ZF}} \qquad (2.3)$$

and that Z may be obtained from ZF by selecting cubes which do not have any f=1 entries. This means that only ZF need be derived by processing the entire network and all other necessary equations may be obtained from this single pass on the input description.

ZF1, fig. 2.1e, describes all fault and input conditions which produce a zero to one error in the output and similarly ZF0 (not shown) describes all conditions for which a one to zero fault occurs. The total fault behavior for the original network is given by

$$ZFT = ZF1 \cap ZF0 \qquad (2.4)$$

ZFT, fig. 3.1b, describes the multiple as well as the single fault conditions for the original network. In addition it is worth noting that all fault sets which are indistinguishable from the good output Z are removed in the process of forming ZFT.

If consideration is to be given to single faults only then ZFT may be modified by selecting only single fault forms of f to give ZFTS. This can be accomplished by

$$ZFTS = ZFT \cap FSF \qquad (2.5)$$

intersecting a cubical representation of single fault values, as in FSF of fig. 2.2, with ZFT as in equation (2.5).

$$\begin{array}{cc} & \begin{array}{ccccccccc} f_6 & f_5 & f_4 & f_3 & f_2 & f_1 & A & B & C \end{array} \\ FSF \Rightarrow & \left\{ \begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & D & D & D \\ 0 & 1 & 0 & 0 & 0 & 0 & D & D & D \\ 0 & 0 & 1 & 0 & 0 & 0 & D & D & D \\ 0 & 0 & 0 & 1 & 0 & 0 & D & D & D \\ 0 & 0 & 0 & 0 & 1 & 0 & D & D & D \\ 0 & 0 & 0 & 0 & 0 & 1 & D & D & D \end{array} \right\} \end{array}$$

Figure 2.2. Cube for Selecting Single Faults

At this point it is important to point out the correspondence between the ZFT, ZFTS equations and the detection matrix G_D, described by Kautz [5]. The nature of the derivation of these equations is such that there is a min term present for each one entry in the detection matrix, G_D. The equation under any particular fault condition f^i, given by $ZFT(f^i,x)$ or $ZFTS(f^i,x)$ corresponds to a column in G_D. Similarly $ZFT(f,x^i)$ or $ZFTS(f,x^i)$ correspond to the row function for row i.

If the objective is to derive tests which isolate to a particular fault class then a new equation can be found. Relations (2.6) would describe the G_I matrix for single faults and (2.7) for multiple faults. With the proper encoding of fault variables

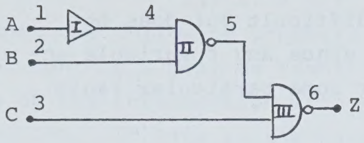
$$ZFTSI \Rightarrow ZFTS(f^i,x) \oplus ZFTS(f^j,x) \cup ZFTS \qquad (2.6)$$

$$\forall\ i,\ j \quad i \neq j$$

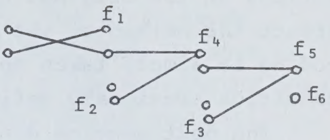
$$ZFTI \Rightarrow ZFT(f^i,x) \oplus ZFT(f^j,x) \cup ZFT \qquad (2.7)$$

$$\forall\ i,\ j \quad i \neq j$$

There is in general the necessity to expand the number of fault variables to encode the various pairwise combinations which result.



Original Network [9]
Figure 2.1a



Connected sets
Figure 2.1b

$$\text{I}\quad \begin{array}{c} \begin{array}{ccc} 1 & f_1 & f_4 \end{array} \\ \left\{\begin{array}{ccc} D & 1 & 0 \\ 0 & D & 0 \end{array}\right\} \end{array} \qquad \text{II}\quad \begin{array}{c} \begin{array}{cccccc} 2 & 4 & f_1 & f_2 & f_4 & f_5 \end{array} \\ \left\{\begin{array}{cccccc} 0 & D & D & 0 & D & 0 \\ D & 0 & 0 & D & D & 0 \\ D & D & D & D & 1 & 0 \end{array}\right\} \end{array}$$

$$\text{III}\quad \begin{array}{c} \begin{array}{cccccc} 3 & 5 & f_3 & f_4 & f_5 & f_6 \end{array} \\ \left\{\begin{array}{cccccc} 0 & D & D & 0 & D & 0 \\ D & 0 & 0 & D & D & 0 \\ D & D & D & D & 1 & 0 \end{array}\right\} \end{array}$$

Cubical Node Descriptions
Figure 2.1c

$$\begin{aligned} ZF = {} & \overline{f}_6\ f_5\ + \\ & \overline{f}_6\ \overline{f}_3\ \overline{C}\ + \\ & \overline{f}_6\ \overline{f}_4\ f_2\ f_1\ + \\ & \overline{f}_6\ \overline{f}_4\ f_2\ \overline{A}\ + \\ & \overline{f}_6\ \overline{f}_4\ f_1\ B\ + \\ & \overline{f}_6\ \overline{f}_4\ \overline{A}\ B \end{aligned} \qquad \begin{array}{c} \begin{array}{ccccccccc} f_6 & f_5 & f_4 & f_3 & f_2 & f_1 & A & B & C \end{array} \\ \left\{\begin{array}{ccccccccc} 0 & 1 & D & D & D & D & D & D & D \\ 0 & D & D & 0 & D & D & D & D & 0 \\ 0 & D & 0 & D & 1 & 1 & D & D & D \\ 0 & D & 0 & D & 1 & D & 0 & D & D \\ 0 & D & 0 & D & D & 1 & D & 1 & D \\ 0 & D & 0 & D & D & D & 0 & 1 & D \end{array}\right\} \end{array}$$

Fault Equation and Cube Form
Figure 2.1d

$$\begin{aligned} ZF1 = {} & \overline{f}_6\ f_5\ (AC + \overline{B}C)\ = \\ & \overline{f}_6\ \overline{f}_4\ f_2\ f_1\ (AC + \overline{B}C) + \\ & \overline{f}_6\ \overline{f}_4\ f_2\ (\overline{A}\ \overline{B}\ C)\ + \\ & \overline{f}_6\ \overline{f}_4\ f_1\ (A\ B\ C) \end{aligned} \qquad \begin{array}{c} \begin{array}{ccccccccc} f_6 & f_5 & f_4 & f_3 & f_2 & f_1 & A & B & C \end{array} \\ \left\{\begin{array}{ccccccccc} 0 & 1 & D & D & D & D & 1 & D & 1 \\ 0 & 1 & D & D & D & D & D & 0 & 1 \\ 0 & D & 0 & D & 1 & 1 & 1 & D & 1 \\ 0 & D & 0 & D & 1 & 1 & D & 0 & 1 \\ 0 & D & 0 & D & 1 & D & 0 & 0 & 1 \\ 0 & D & 0 & D & D & 1 & 1 & 1 & 1 \end{array}\right\} \end{array}$$

0-1 Fault Equation and Cube Form
Figure 2.1e

This required encoding tends to make the correspondence to conditions in the original network more difficult but does not affect the method of the next section since any f variable encoding is simply taken to be a tag for some particular fault condition previously defined.

The next section describes how a test set can be derived using the equations described here. The nature of the equation used, *i.e.* ZFT, ZFTS, ZFTI or ZFTSI, will determine the applicability of the test set to detect or isolate single or multiple faults.

3. TEST GENERATION FROM A FAULT EQUATION

Once the desired equation has been derived then the selection of tests is analogous to the selection of rows from the appropriate fault matrix. The approach taken here will be to use the algebraic properties of the cubical equation to accomplish this task. Properties of cubical equations are covered in detail elsewhere [3][6].

The input equation G(f,x) may be either ZFT, ZFTS, ZFTI or ZFTSI and is shown expanded as

$$G(f,x) = f^0G(0,x) + f^1G(1,x) + \ldots f^{2^m-1}G(2^m-1,x) \qquad (3.1)$$

where the min terms of f and x vectors are given by

$$f^{\cdot} = f^0, f^1, \ldots, f^{2^m-1} \qquad (3.2)$$

$$x^{\cdot} = x^0, x^1, \ldots, x^{2^n-1} \qquad (3.3)$$

Each min term of $f^{\cdot}$ represents some single or multiple fault condition(s) in a column of either a detection, G_D, or isolation, G_I, matrix. Each min term of $x^{\cdot}$ represents an input and possible test for the network.

The objective of test selection, independent of the nature of G, is to select a test function T, as in equation 3.4, so that the relations

$$\begin{aligned} &T = t^1 + t^2 \ldots t^j \qquad (3.4) \\ &t^i \in x^{\cdot} \quad i = 1,2,\ldots j \end{aligned}$$

of (3.5) hold. In equation 3.5 i is used for f^i, etc. That is to say each fault

$$\begin{aligned} &G(0,x) \cap T \neq \emptyset \\ &G(1,x) \cap T \neq \emptyset \\ &\vdots \qquad\qquad\qquad\qquad\qquad (3.5) \\ &G(2^m-1,x) \cap T \neq \emptyset \qquad \text{for all } G(i,x) \neq \emptyset \end{aligned}$$

column present in G must have at least one term which intersects with the test function. If the set of intersections of (3.5) is rewritten as cubical equation (3.6)

$$G(f,x) \cap T(x) = H(f,x) \tag{3.6}$$

then the following property can be noted. The f min terms present in H must be identical with the f min terms present in G or the intersection with the test function has eliminated some column. If this is the case, then at least one of the relations of (3.5) is not satisfied.

The following definition allows further use of the cubical operators. H(f,D) is defined

$$(H(f,x))_{DX} = H(f,D) \tag{3.7}$$

as the H equation of (3.6) with all x variables as "don't cares" or D's. The cubical sharp operation can be used to remove the f min terms of H from those in G and the resulting cube should be empty.

$$G(f,x) \# H(f,D) = \emptyset \tag{3.8}$$

This expression says in effect that the f terms of H are removed from G. If the result is that no terms remain then the two f sets were equal since there can never be any new f terms in H which were not in G.

The test then for an acceptable test function T is given by

$$G(f,x) \# (G(f,x) \cap T(D,x))_{DX} = \emptyset \tag{3.9}$$

Any T for which equation (3.9) holds represents an acceptable test set.

There are 2^{2^n} possible selections of subsets of $x^{\cdot}$ and hence T(x). If the possible T are selected as all sets with one element, all with two, etc., then the first acceptable T represents the smallest possible set.

The example of section 2 is shown completed in fig. 3.1 for a particular T. The resulting H(f,D) of fig. 3.1c is sharped with fig. 3.1b. It can be seen that all entries in ZFT (fig. 3.1b) are included in some entry of 3.1c and hence for this example the T shown is acceptable and involves four test inputs.

$$T = T^0 + T^1 + T^3 + T^7 \quad \begin{array}{c} \begin{array}{ccccccccc} f_6 & f_5 & f_4 & f_3 & f_2 & f_1 & A & B & C \end{array} \\ \left\{ \begin{array}{ccccccccc} D & D & D & D & D & D & 0 & 0 & 0 \\ D & D & D & D & D & D & 0 & 0 & 1 \\ D & D & D & D & D & D & 0 & 1 & 1 \\ D & D & D & D & D & D & 1 & 1 & 1 \end{array} \right\} \end{array}$$

One Possible Test Function

Figure 3.1a

$$\mathrm{ZFT}\quad \begin{array}{c} \begin{array}{ccccccccc} f_6 & f_5 & f_4 & f_3 & f_2 & f_1 & A & B & C \end{array} \\ \left\{ \begin{array}{ccccccccc} 0 & 1 & D & D & D & D & 1 & D & 1 \\ 0 & 1 & D & D & D & D & D & 0 & 1 \\ 0 & D & 0 & D & 1 & 1 & 1 & D & 1 \\ 0 & D & 0 & D & 1 & 1 & D & 0 & 1 \\ 0 & D & 0 & D & 1 & D & 0 & 0 & 1 \\ 0 & D & 0 & D & D & 1 & 1 & 1 & 1 \\ 1 & D & D & D & D & D & 0 & 1 & D \\ 1 & D & D & D & D & D & D & D & 0 \\ D & 0 & 1 & 1 & D & D & 0 & 1 & D \\ D & 0 & 1 & 1 & D & D & D & D & 0 \\ D & 0 & D & 1 & 0 & D & D & 0 & 0 \\ D & 0 & D & 1 & D & 0 & 1 & D & 0 \\ D & 0 & 1 & D & D & D & 0 & 1 & 1 \end{array} \right\} \end{array}$$

Multiple Fault Detection Equation

Figure 3.1b

$$(\mathrm{ZFT} \cap \mathrm{T})_{\mathrm{DX}}\quad \begin{array}{c} \begin{array}{ccccccccc} f_6 & f_5 & f_4 & f_3 & f_2 & f_1 & A & B & C \end{array} \\ \left\{ \begin{array}{ccccccccc} 0 & 1 & D & D & D & D & D & D & D \\ 0 & D & 0 & D & 1 & 1 & D & D & D \\ 0 & D & 0 & D & 1 & D & D & D & D \\ 0 & D & 0 & D & D & 1 & D & D & D \\ 1 & D & D & D & D & D & D & D & D \\ D & 0 & 1 & 1 & D & D & D & D & D \\ D & 0 & D & 1 & 0 & D & D & D & D \\ D & 0 & 1 & D & D & D & D & D & D \end{array} \right\} \end{array}$$

H(f,D) of equation (3.8)

Figure 3.1c

4. CONCLUSIONS

The use of the fault equation as a model of behavior seems to offer several advantages. The most important property is that the form of the equation need not be maintained to preserve the network structure. In using the equation to derive a test set, only a single pass is required to derive all necessary information from the original network. Since the equation can be carried in disjunctive form, symbol manipulation can be avoided.

The number of variables required does grow rapidly with the size of the network and selection of isolation tests does require a large amount of data encoding. This however must be compared to previous methods in which the number of iteration precluded any attempt at considering multiple faults. Feasibility of the approach has been shown for a small number of variables (32 max) using Fortran IV.

REFERENCES

1. D. B. Armstrong, "On Finding a Nearly Minimal Set of Fault Detection Tests for Combinational Logic Nets," IEEE Trans. on Electronic Computers, vol. EC-15, no. 1, pp. 66-73, February, 1966.

2. H. Y. Chang, "An Algorithm for Selecting an Optimum Set of Diagnostic Tests," IEEE Trans. on Electronic Computers, vol. EC-14, no. 5, pp. 706-711, October, 1965.

3. D. L. Dietmeyer, Logic Design of Digital Systems, Allyn-Bacon, Boston, Massachusetts, 1971.

4. J. W. Gault, "Processing for Fault Behavior in Digital Networks", Research Report, North Carolina State University, Department of Electrical Engineering, Raleigh, N. C., August 1971.

5. W. H. Kautz, "Fault Testing and Diagnosis in Combinational Circuits", IEEE Trans. on Computers, vol. C-17, no. 4, pp. 352-366, April, 1968.

6. R. E. Miller, Switching Theory, vol. 1, Wiley & Sons, New York, 1965.

7. J. F. Poage, "Derivation of Optimum Tests to Detect Faults in Combinational Circuits", Mathematical Theory of Automata, Polytechnic Press, Brooklyn, N. Y., 1963.

8. J. P. Roth, W. G. Bouricuis, and P. R. Schneider, "Programmed Algorithms to Compute Tests to Detect and Distinguish Between Failures in Logic Circuits", IEEE Trans. on Electronic Computers, vol. EC-16, no. 5, pp. 567-580, October 1965.

9. D. R. Shertz, "On the Representation of Digital Faults", Coordinated Science Laboratory Report R-418, University of Illinois, May, 1969.

CIRCUIT STRUCTURE AND SWITCHING FUNCTION VERIFICATION

MIN-WEN DU AND C. DENNIS WEISS
Computer Science, and Electrical Engineering
Johns Hopkins University, Baltimore, Maryland 21218

ABSTRACT

A new approach is presented for the design of multiple fault detection tests in which the structure of a combinational circuit is used to reduce the number of input combinations required. The structure is defined by the interconnection of the basic elements, each of arbitrary complexity. The fault model assumes that the functions realized by the basic elements may undergo any deviation whatsoever, but that the circuit structure is fault free. Thus, arbitrary combinations of multiple faults within one or more basic elements are included in the model. Decomposition theory can be used to verify that a set of input combinations is a multiple fault detection test set under this model. A process called expansion will be introduced to simplify this task. A well-defined procedure is given for deriving a suitable test set which for some circuits is minimal or near minimal. It will yield a multiple fault detection test of length less than 2^n for any circuit with a nontrivial nondisjoint decomposition defined by a basic-element partition. Higher order basic-element partitions are introduced as a generalization. An upper bound is given on the length of a multiple fault detection test for any circuit with a given structure, independent of the function realized on the structure. The bound is tighter when function information is also used.

I INTRODUCTION

The designing of tests for multiple faults in combinational circuits has proven to be extremely difficult, primarily because most approaches require the consideration of each combination of possible single faults, usually assumed to be of the stuck-at-1 or stuck-at-0 variety. If there are n lines in a circuit, there are 3^n-1 possible multiple stuck faults. Although these may be considered in equivalence classes [1], the number of such classes may still be quite large.

The most common approaches to the design of fault detection tests involve either simulation [2] or employ the basic path sensitization idea [3,4], sometimes with the use of the Boolean Difference [5]. In the present study, a quite different approach is possible, largely because of the more general fault model used. Rather than viewing a multiple fault as a combination of line-stuck faults, we consider it to be an arbitrary transformation of the function realized in certain basic elements of the circuit. This model has been explored by Kautz [6], Thurber [7], and Hornburkle and Spann [8], ordinarily under the assumption that only a single basic element is faulty at any time. (In [6,7] only array circuits are considered). The more general model which permits any number of basic elements to be faulty has also been considered for a limited class of networks [9].

A multiple fault detection test for a circuit with this fault model will be called a function verification test set. This more precisely characterizes the property of a test set and emphasize that the model will handle faults which are not only due to device failure but to incorrect assembly, part selection or part labeling.

II FORMULATION OF THE FUNCTION VERIFICATION PROBLEM

A combinational circuit C will be viewed as an interconnection of basic switching elements, $E_1, \ldots., E_k$. Each is either an individual gate or itself a combinational circuit with inputs and one output*. C will be assumed to be free of feedback loops and to have only one circuit output, f_C. Each input of a basic element is driven by either a primary input variable from the set $I=\{I_1,\ldots.,I_n\}$, or by the output of another basic element.

Definition 1 The variation of a basic element E_i, $\mathrm{Var}(E_i)$, is the set of possible faulty functions which E_i may realize. Here we let $\mathrm{Var}(E_i)=\{g|g\colon \{0\ 1\}^{m_i}\rightarrow\{0,1\}\}$, where m_i is the number of inputs to E_i. Thus, $\mathrm{Var}(E_i)$ is the set of all m_i-place switching functions. $\mathrm{Var}(C)=[\mathrm{Var}(E_1)\ldots.,\mathrm{Var}(E_k)]$ will denote the variation of C, where $E_1,\ldots.,E_k$ are all the basic elements in circuit C.

Rather than selecting functions for $\mathrm{Var}(E_i)$ which could be obtained by assuming all possible combinations of single or multiple stuck faults on lines within E_i, we have chosen a more encompassing model which includes these as well as more general multiple faults due perhaps to incorrect part selection on wiring within any basic element E_i.

Let f_C denote the correct output function realized by C, and let f_g denote an output function obtained if the ith basic element in C realizes the ith component of $g=(g_1,\ldots.,g_k)$. Then the set of faulty functions realizable by C is $F_C = \{f_g|g\varepsilon\mathrm{Var}(C)\}$. Note that F_C does not depend on the particular functions realized by the basic elements in C but only on the topological structure of the circuit.

---

*A k- output element can be replaced by k parallel single-output basic elements, each with the original inputs.

Definition 2 (a) A set of input combinations $S \subseteq 2^{[I]}$ is said to be able to *detect* fault g iff $f_g(S) \neq f_C(S)$.

(b) S is a *function verification test set* (FVTS) for C iff $\forall f_g \epsilon F_C, \; f_g(S) = f_C(S) \rightarrow f_g = f_C$.

(c) S is a *minimal* FVTS of C iff there is no other FVTS which is smaller. $2^{[I]}$ will denote the Cartesian product $\{0,1\}^n$ where n is the number of variables in I.

III BASIC-ELEMENT PARTITIONS AND SETS OF COMPARABLE CONFIGURATIONS

Certain partitions of the input variable set will be of interest, each associated with a basic element of the circuit. Each partition will define a classical decomposition of f_C [10,11,12,13], which is, in general, nondisjoint. The existence of these decompositions imply certain restrictions on the function realized by the circuit which can be expressed in terms of the relation of certain subfunctions of f_C called *comparable configurations*. These, in turn, will be used to derive an FVTS.

Definition 3 For any basic element E_j in C, a *basic-element partition* P_j is defined by $P_j = [X|Y\|Z]$ where $Z = IN(E_j) \cap OUT(E_j)$, $X = IN(E_j) - Z$, and $Y = OUT(E_j) - Z$, with $IN(E_j) = \{x\colon x \epsilon I$ and there exists a path from some occurrence of a primary input variable x to $E_j\}$, $OUT(E_j) = \{x\colon x \epsilon I$ and there exists a path from some occurrence of a primary input variable x to the circuit output which does not pass through $E_j\}$. If $|X| < 2$ or $Y = \emptyset$, then P_j is called a *trivial* partition. If $Z = \emptyset$, then write $P_j = [X|Y]$.

The significance of the blocks in P_j is seen by noting that P_j defines a decomposition of the circuit output function $f_C(I)$ as

$$f_C(I) = f_j(h_j(X,Z),Y,Z) \tag{1}$$

where h_j is the output function of basic element E_j expressed in terms of primary input variables, and f_j is the circuit output function expressed in terms of the output of E_j and the primary input variables. Clearly, variables in X control the circuit output only by affecting the value of h_j; and variables in Y have no effect on the value of h_j. Note that if Z is empty, (1) is called a disjoint decomposition of f_C; otherwise, it is a nondisjoint decomposition.

Definition 4 Let $P_j = [X|Y\|Z]$ be a nontrivial basic element partition of a circuit realizing f, and x,y,z denote specific binary assignments to the variables in X,Y and Z, respectively. Then a subfunction $f_C(X,y,z)$ will be called a *configuration* (CFG) of f_C; and $\{f_C(X,y_i,z)\colon y_i \epsilon 2^{[Y]}\}$ a complete set of *comparable configurations* associated with P_j. Any two configurations in the set are said to be *comparable*.

Note that there are $|2^{[Z]}|$ complete sets of comparable configurations associated with the partition $P_j = [X|Y\|Z]$.

The following theorem is a generalization of the basic idea in Ashenhurst [10].

Theorem 1 Every subfunction in a complete set of comparable configurations associated with a single basic-element partition is one of the functions $0, 1, \psi(X), \psi'(X)$, where $0(1)$ denotes the identically zero (one) function, and $\psi(X)$ is some switching function which depends nonvacuously on at least one variable in X.

Definition 5 $f_C(X,y,z)$ will be referred to as a 0-, 1-, or ψ^*-CFG, where ψ^* denotes either $\psi(X)$ or $\psi'(X)$.

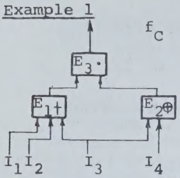
Proof of Theorem 1:

If $P_j = [X|Y\|Z]$ is a basic-element partition, then $f_C(X,y,z) = f_j(h_j(X,z),y,z) \epsilon \{0,1,\psi(X),\psi'(X)\}$ where $\psi(X) = h_j(X,z)$.

C is a *tree circuit*, if every primary input variable and basic element output goes to only one basic element input, and there are no loops. For tree circuits, $Z = \emptyset$ in any basic-element partition P_j. For such a circuit, decomposition theory provides a simple description of the necessary and sufficient conditions on a function in order that it be realizable on a given tree-like structure.

Theorem 2 [Ashenhurst, Karp [10,12]] If C is a tree-like structure, then $f \epsilon F_C$ if and only if for every $P_j = [X|Y\|\emptyset]$, $\{f(X,y_i)\colon y_i \epsilon 2^{[Y]}\} \subseteq \{0,1,\psi,\psi'\}$ for some function $\psi(X)$.

Example 1

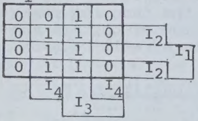


$IN(E_1) = \{I_1,I_2,I_3\}$
$IN(E_2) = \{I_3,I_4\}$
$IN(E_3) = \{I_1,I_2,I_3,I_4\}$

$OUT(E_1) = \{I_3,I_4\}$
$OUT(E_2) = \{I_1,I_2,I_3\}$
$OUT(E_3) =$ empty

Thus, $P_1 = [I_1I_2|I_4\|I_3]$, $P_2 = [I_4|I_1I_2\|I_3]$, $P_3 = [I_1I_2I_3I_4|$ empty $\|$ empty$]$.
Only P_1 is a nontrivial basic-element partition.

$f_C =$



I_2
I_1
I_2
I_4
I_4
I_3

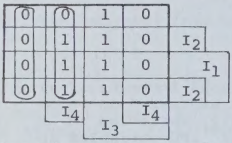
0	0	1	0
0	1	1	0
0	1	1	0
0	1	1	0

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

reference map with decimal equivalents of input combinations

Corresponding to P_1, we have the following two complete sets of comparable CFGs. Each CFG is encircled.

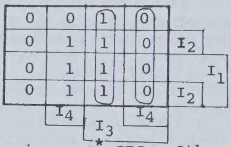
(a)



I_2
I_1
I_2
I_4
I_4
I_3

0	0	1	0
0	1	1	0
0	1	1	0
0	1	1	0

(b)



I_2
I_1
I_2
I_4
I_4
I_3
*

0	0	1	0
0	1	1	0
0	1	1	0
0	1	1	0

Of these four CFGs, only $f_C(I_1I_2;I_4=1;I_3=0)$ is a ψ^*-CFG. Others are either 1-CFGs or 0-CFGs.

IV ON DETERMINING WHETHER A SET IS AN FVTS FOR A GIVEN CIRCUIT

Let $S\subseteq 2^{[I]}$, where I is the input variable set of a circuit C. Then the deviation of f_C on S, $Dev(f_C,S)$, is the set of all other functions which agree with f_C on the domain S.

$$Dev(f_C,S) = \{f|f\colon 2^{[I]}\rightarrow\{0,1\},\quad f(S) = f_C(S), \text{ and } f\neq f_C\}$$

We will also let $Res(f,S)$ denote the restriction of f to S. Such a partial function will be considered as a set of ordered pairs.

In principle, it is a simple matter to determine whether S is an FVTS for a circuit C. For each $f\in Dev(f_C,S)$, one determines whether f can be realized on the given circuit topology, a task for which straightforward though tedious procedures exist [11,13]. If C is a tree-circuit, Theorem 2 can be used for testing each f. However, for large circuits the procedure would still be lengthy. A simpler procedure can be formulated in which S is expanded to a larger set T so that $Dev(f_C,T)$ is reduced. If T becomes the full domain $2^{[I]}$, then S is an FVTS of C.

Definition 6 Let $P_j = [X|Y\|Z]$ be a basic-element partition of a circuit C and $T\subseteq 2^{[I]}$. Then we say that $f_C(X\ y\ z)$ has been verified by T to be a ψ^*-CFG if there exist $(x_1yz),(x_2yz)\in T$ and $f_C(x_1yz)\neq f_C(x_2yz)$.

Definition 7 Let $P = \{P_1,\ldots,P_k\}$ be the set of basic-element partitions of a circuit C, and $S\subseteq 2^{[I]}$. Then the expansion of $Res(f_C,S)$, $Ex(f_C,S)$, is a partial function with domain T, defined recursively as

(a) $Res(f_C,S)\subseteq Ex(f_C,S)$

(b) If for some $P_j = [X|Y\|Z]$, $f_C(Xy_\alpha z)$ and $f_C(Xy_\beta z)$ have been verified by T to be ψ^*-CFGs, and there exist $x,x'\in X$ such that $(xy_\alpha z),(x'y_\alpha z)$ and $(xy_\beta z)\in T$, then

$$((x'y_\beta z),\ f_C(x'y_\beta z))\in Ex(f_C,S).$$

We will say that this pair can be derived by a ψ-ψ implication. $f_C(x'y_\beta z)$ equals (is the complement of) $f_C(x'y_\alpha z)$ if $f_C(xy_\beta z)$ equals (is the complement of) $f_C(xy_\alpha z)$.

(c) If for some $P_j = [X|Y\|Z]$, $f_C(Xy_\alpha z)$ has been verified by T to be a ψ^*-CFG, and there exists x_1,x_2,y_β such that $(x_1y_\alpha z)$, $(x_2y_\alpha z)$, $(x_1y_\beta z)$, $(x_2y_\beta z)\in T$ and $f_C(x_1y_\alpha z)\neq f_C(x_2y_\alpha z)$, $f_C(x_1y_\beta z) = f_C(x_2y_\beta z) = k$, then for all $x\in X$,

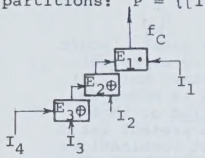
$$((xy_\beta z),\ f_C(xy_\beta z))\in Ex(f_C;\ S).$$

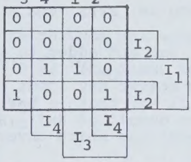
Each of these pairs can be derived by a ψ-0 (for k=0) or ψ-1 (for k=1) implication.

$f_C(xy_\beta z) = k$ for all $x \epsilon X$.

Operationally, it is straightforward to compute $Ex(f_C,S)$. One begins with $\{(i,f_C(i)): i \epsilon S\}$ and, using all non-trivial basic-element partitions of C, derives as many new values of f_C as possible using the implications in Definition 7. Iteratively, newly generated pairs are combined with the initial pairs to generate still others until no new pairs can be derived from any partition and, by definition, $Ex(f_C,S)$ has been obtained.

Example 2 Consider the following circuit with the set of nontrivial basic-element partitions: $P = \{[I_2I_3I_4 | I_1],[I_3I_4 | I_1I_2]\}$





0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

Let $S = \{0,1,8,9,12,14\}$. $Ex(f_C,S)$ is obtained step by step as follows:

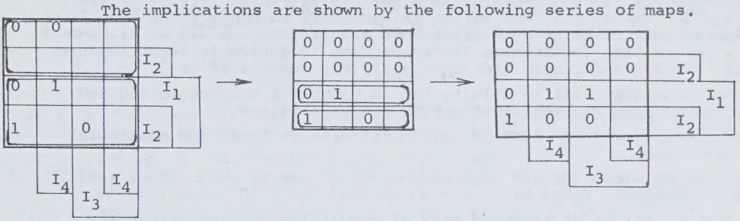
(a) $\{(0,0),(1,0),(8,0),(9,1),(12,1),(14,0)\} \subseteq Ex(f_C,S)$.

(b) $\{(2,0),(3,0),(4,0),(4,0),(6,0),(7,0)\} \subseteq Ex(f_C,S)$
by ψ-0 implications between the CFGs $f_C(I_2I_3I_4; I_1 = 1)$ and $f_C(I_2I_3I_4; I_1 = 0)$.

(c) $\{(10,1),(13,0)\} \subseteq Ex(f_C,S)$ by ψ-ψ implications between the CFGs $f_C(I_3I_4; I_1=1, I_2=0)$ and $f_C(I_3I_4; I_1=1,I_2=1)$.

(d) No further implications are available so that
$T = \{0,1,2,3,4,5,6,7,8,9,10,12,13,14\}$.

The implications are shown by the following series of maps.

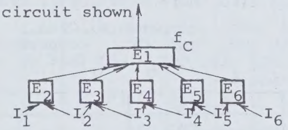


Res(f_C,S) after Step (b) $Ex(f_C,S)$

Theorem 3 A sufficient condition that $S \subseteq 2^{[I]}$ be an FVTS for a circuit C is that the domain T of $Ex(f_C,S)$ is the full set $2^{[I]}$.

Proof: Direct consequence of Theorem 1. It is not, in general necessary that the domain of $Ex(f_C,S)$ be the full set $2^{[I]}$ if S is an FVTS, as shown by the following example.

Example 3 Consider the circuit shown



There are $2^{2^6} = 2^{64}$ distinct functions from $2^{[I]}$ to $\{0,1\}$, but the above structure can realize at most $2^{2^5} \cdot (2^{2^2})^5 = 2^{52}$ functions, the total number of distinct ways of assigning functions to the basic elements. Hence, there exists $f \epsilon F_C$. But then there exist $f_1 \epsilon F_C$, $f_2 \notin F_C$ which differ for only one domain element, u . This is because there is a chain $g_1, \ldots, g_{n+1}$ such that $g_1 = f_C$, $g_{n+1} = f$, and g_i differs from g_{i+1} for only one domain element, $k = 1,\ldots,n$. But then $f_1(2^{[I]}-\{u\}) = f_2(2^{[I]}-\{u\})$ and $f_1(u) \neq f_2(u)$. Thus $S = 2^{[I]}-\{u\}$ is an FVTS of f_1 because the structure can realize f_1 since $f_1 \epsilon F_C$. On the other hand, the domain of $Ex(f_1,S)$ is S since the circuit above has no nontrivial basic-element partitions. Therefore, the converse of Theorem 3 is not always true.

We conjecture that the converse of Theorem 3 is true for tree-circuits. No proof has yet been found, but efforts at constructing counter-examples have failed.

In particular, if the domain of $Ex(f_C,S)$ is not $2^{[I]}$, where C is a tree-circuit, then the following construction appears to always yield a function $f \epsilon F_C$ $f \neq f_C$. Let the partial function $Ex(f_C,S)$ have domain $T_1 \subset 2^{[I]}$. Then select some $u_1 \notin T_1$ for which $Ex(f_C, S \cup \{u_1\})$ has domain $T_2 \subset 2^{[I]}$. Continue until $Ex(f_C, S \cup \{u_1,\ldots,u_{i-1}\})$ has domain $T_i \subset 2^{[I]}$ but for every $u \notin T_i$, $Ex(f_C, S \cup \{u_1,\ldots,u_{i-1},u\})$ has domain $2^{[I]}$.

Then let $f = (T_i, f_C(T_i)) \cup (2^{[I]}-T_i, \overline{f}_C(2^{[I]}-T_i))$. In all cases investigated to date, f does not violate Theorem 2 so that S is not an FVTS of the tree-circuit C.

V A PROCEDURE TO FIND AN FVTS FOR A CIRCUIT

A simple procedure is presented here for finding an FVTS for a given circuit. Input combinations are picked up one by one and added to an originally empty set S. The process stops when the domain of $Ex(f_C,S)$ is $2^{[I]}$. Of course, $Ex(f_C,S)$ is recomputed (the expansion process) whenever any input combination is added to S. At any step in the procedure, we will say that u has been verified or that $f_C(u)$ has been verified if u is in the domain of $Ex(f_C,S)$ for the present set S. Two basic rules will be useful for determining which unverified input combinations to add to S.

(a) To apply ψ-0 or ψ-1 implications. Consider two comparable CFGs, $f_C(Xy_\alpha z)$ and $f_C(Xy_\beta z)$, a ψ^*-CFG, and a 0- (or 1-) CFG, respectively. If $f_C(Xy_\alpha z)$ has been verified to be a ψ^*-CFG(Definition 6), then two input combinations in $(Xy_\beta z)$ are sufficient to verify that $f_C(Xy_\beta z)$ is a 0- (or 1-) CFG, namely $(x_1 y_\beta z)$, $(x_2 y_\beta z)$ where x_1 and x_2 are such that $f_C(x_1 y_\alpha z) \neq f_C(x_2 y_\alpha z)$.

(b) To apply ψ-ψ implications. Consider two comparable ψ^*-CFGs, $f_C(Xy_\alpha z)$ and $f_C(Xy_\beta z)$. If $f_C(Xy_\alpha z)$ has been verified to be a ψ^*-CFG by verification of $f_C(\hat{X}y_\alpha z)$ for some $\hat{X} \subset 2^{[I]}$, and if no $f_C(Xy_\beta z)$ has been verified, it is desirable to verify that $f_C(Xy_\beta z)$ is a ψ^*-CFG using a pair $(x_1 y_\beta z)$ with $x_1 \epsilon 2^{[I]}-\hat{X}$, and $(x_2 y_\beta z)$ with $x_2 \epsilon \hat{X}$. The second combination is necessary to connect the two ψ^*-CFGs together so that it is known whether both are ψ-CFGs, or one is ψ, one ψ'. Now a ψ-ψ implication (Definition 7) can be used to derive at least two new function values, namely $f_C(x_1 y_\alpha z)$ and $f_C(x_i y_\beta z)$ for some $x_i \epsilon \hat{X}-\{x_2\}$.

Some other techniques will be illustrated in Example 4 for connecting and verifying ψ^*-CFGs so that an even shorter FVTS can be obtained.

Procedure 1 For constructing an FVTS for a circuit with at least one nontrivial basic-element partition.

In the following steps, we add combinations to S one by one. After each addition to S, recompute $E_x(f_C,S)$.

Step 1 Order the collection of complete sets of comparable configurations (by size of configuration in the set, for example).

Step 2 Verify one ψ^*-CFG in each complete set of comparable CFGs (see Definition 6). If CFG_1, CFG_2 are ψ^*-CFGs in the comparable sets G_1 and G_2 respectively, then if CFG_1 has been verified, and $CFG_2 \subset CFG_1$, we will choose CFG_2 to be the ψ^*-CFG in G_2 which is to be verified.*

Step 3 Verify the 0- and 1-CFGs (with the selection rule stated below) which are contained in the ψ^*-CFGs verified in Step 2, by using rule (a).

The following selection rule applies if $CFG_2 \subset CFG_1$ where CFG_1, CFG_2 are both ψ^*-CFGs verified in Step 2, and also there exist CFG_3, CFG_4, CFG_5 such that

(i) $CFG_3 \subset CFG_1$ with CFG_2 and CFG_3 comparable CFGs.
(ii) $CFG_4 \subset CFG_2$, $CFG_5 \subset CFG_3$, with CFG_4 and CFG_5 comparable 0- or 1-CFGs.
(iii) The relative position of CFG_4 in CFG_2 and that of CFG_5 in CFG_3 are the same, as shown in Fig. 1

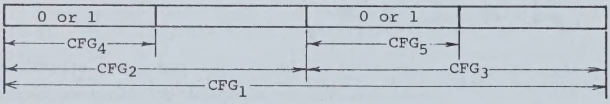


Fig. 1. Relation between CFGs for the selection rule in Step 3.

Then verify CFG_4 but not CFG_5, because in the next step we will connect CFG_2 and CFG_3 so that CFG_5 is then implied by CFG_4.

Step 4 Verify and connect all ψ^*-CFGs by using rule (b).

Step 5 Verify all the remaining 0-CFGs and 1-CFGs which are comparable to other ψ^*-CFGs, by using rule (a).

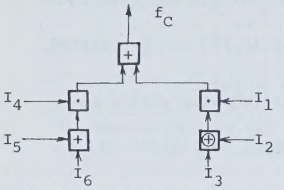
---

* It can be shown that if CFG_2 is contained in both CFG_1 and $CFG_{1'}$, then either $CFG_1 \subseteq CFG_{1'}$, or $CFG_{1'} \subseteq CFG_1$, so that no more than one ψ^*-CFG in CFG_2 will be verified in Step 2 of Procedure 1.

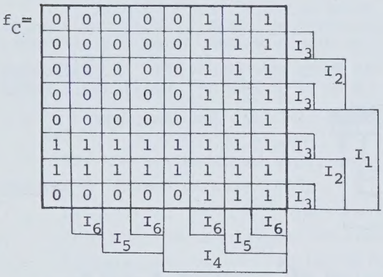
Step 6 Pick up the remaining unverified input combinations one by one until the function value on the full set $2^{[I]}$ is verified.*

Step 7 For every x in S, if the domain of $Ex(f_C, S-x)$ is $2^{[I]}$, delete x from S.

Example 4 Consider this circuit with the set of nontrivial basic-element partitions:



$$P = \{[I_1I_2I_3 | I_4I_5I_6],\ [I_2I_3 | I_1I_4I_5I_6],\ [I_4I_5I_6 | I_1I_2I_3],\ [I_5I_6 | I_1I_2I_3I_4]\}$$



0	0	0	0	0	1	1	1
0	0	0	0	0	1	1	1
0	0	0	0	0	1	1	1
0	0	0	0	0	1	1	1
0	0	0	0	0	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
0	0	0	0	0	1	1	1

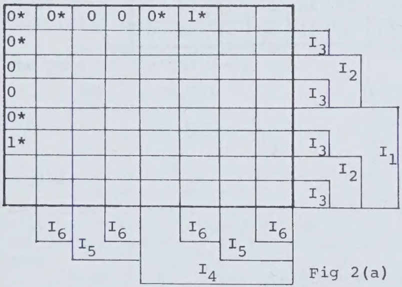
0	1	2	3	4	5	6	7
8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23
24	25	26	27	28	29	30	31
32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47
48	49	50	51	52	53	54	55
56	57	58	59	60	61	62	63

Step 1 All columns, all half columns, all rows and all half rows on the map are sets of comparable CFGs. Call them G_1, G_2, G_3, G_4.

Step 2 Pick up 0, 40 (denoted $\{0,40\} \rightarrow S$); then
$f_C(I_1I_2I_3;\ I_4 = 0,\ I_5 = 0,\ I_6 = 0)$ is verified to be a ψ*-CFG.
$\{32\} \rightarrow S$ verifies that $f_C(I_2I_3;\ I_1 = 1,\ I_4 = 0,\ I_5 = 0,\ I_6 = 0)$ is a ψ*-CFG.
$\{5\} \rightarrow S$ verifies that $f_C(I_4I_5I_6;\ I_1 = 0,\ I_2 = 0,\ I_3 = 0)$ is a ψ*-CFG.
$\{4\} \rightarrow S$ verifies that $f_C(I_5I_6;\ I_1 = 0,\ I_2 = 0,\ I_3 = 0,\ I_4 = 1)$ is a ψ*-CFG.

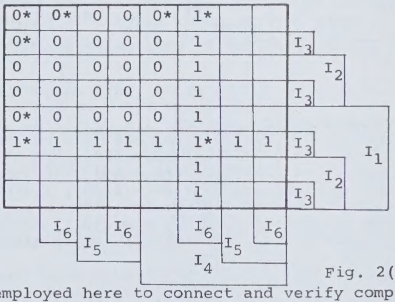
Step 3 $\{8\} \rightarrow S$ verifies that $f_C(I_2I_3;\ I_1 = 0,\ I_4 = 0,\ I_5 = 0,\ I_6 = 0)$ is a 0-CFG.
$\{1\} \rightarrow S$ verifies that $f_C(I_5I_6;\ I_1 = 0,\ I_2 = 0,\ I_3 = 0,\ I_4 = 0)$ is a 0-CFG.

The partial function $Ex(f_C, S)$ at this stage is shown in Fig. 2(a), where S is the set of starred input combinations.



0*	0*	0	0	0*	1*		
0*							
0							
0							
0*							
1*							

Fig 2(a)



0*	0*	0	0	0*	1*		
0*	0	0	0	0	1		
0	0	0	0	0	1		
0	0	0	0	0	1		
0*	0	0	0	0	1		
1*	1	1	1	1	1*	1	1
					1		
					1		

Fig. 2(b)

Step 4 A very useful technique may be employed here to connect and verify comparable ψ*-CFGs. Let $\{45\} \rightarrow S$. We verify that $f_C(I_4I_5I_6;\ I_1=1,\ I_2=0,\ I_3=1)$ is a ψ-CFG. This CFG goes through G_1. It will be called a 1-axis for G_1** (since the function value is one on the axis; 0-axes are similarly defined). Along the axis, connecting and verifying the ψ*-CFGs requires the addition to S of only one input combination for each ψ*-CFG. After {45}→S, Ex(f_C,S) is shown in Fig. 2(b). Let {56}→S, to verify the bottom row to be a ψ*-CFG. Because of the 1-axis $f_C(I_1I_2I_3;\ I_4 = 1,\ I_5 = 0,\ I_6 = 1)$, it is automatically connected to other comparable ψ*-CFGs.

---

* Here, input combinations are selected in an arbitrary order. A 1-level (or higher level) search may be worth considering. A computer program is currently being implemented for the procedure, and experimentation with such local optimizations at Step 6 may be explored.

** An axis may not go through all CFGs in a complete set of comparable CFGs.

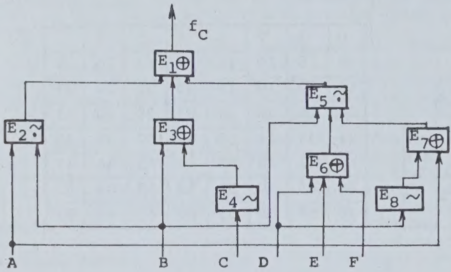
Step 5 {48}, {6}, {7}→S verify all the remaining 1-CFGs.

Step 6 The function values for all input combinations have now been verified because the domain of the expansion is $2^{[I]}$.

Step 7 None of the elements in S = {0,1,4,5,6,7,8,32,40,45,48,56} can be deleted, so that a length 12 FVTS is obtained.

By applying somewhat more complicated rules, a length 10 FVTS S = {0,5,40,45,9,36,60,38,39,52} may be found; the method of [9] also yields an FVTS of length 10, which is believed to be minimum.

Example 5 A nontree-circuit which has nontrivial basic-element partitions is shown below.



E_1,E_3,E_6 and E_7 each realize the odd parity functions of k inputs, E_2 and E_5 realize NAND, and E_4 and E_8 are NOT gates. The set of nontrivial basic-element partitions is P={[EF|ABC‖D],[DEF|C‖AB]}. f_C is given below.

1	1	1	1	1	1	1	1			
0	0	0	0	0	0	0	0	C		
1	0	0	1	0	0	0	0		B	
0	1	1	0	1	1	1	1	C		
1	1	1	1	1	1	1	1			A
0	0	0	0	0	0	0	0	C		
1	1	1	1	1	0	0	1		B	
0	0	0	0	0	1	1	0	C		
	F		F		F		F			
		E				E				
				D						

0	1	2	3	4	5	6	7
8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23
24	25	26	27	28	29	30	31
32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47
48	49	50	51	52	53	54	55
56	57	58	59	60	61	62	63

The sets of comparable CFGs are
$G_1 = \{f_C(DEF;\ C = i:\ i\epsilon\{0,1\};\ AB = 00)\}$
$G_2 = \{f_C(DEF;\ C = i:\ i\epsilon\{0,1\};\ AB = 01)\}$
$G_3 = \{f_C(DEF;\ C = i:\ i\epsilon\{0,1\};\ AB = 10)\}$
$G_4 = \{f_C(DEF;\ C = i:\ i\epsilon\{0,1\};\ AB = 11)\}$
$G_5 = \{f_C(EF;\ ABC = i:\ i\epsilon\ 2^{[ABC]};\ D = 0)\}$
$G_6 = \{f_C(EF;\ ABC = i:\ i\epsilon\ 2^{[ABC]};\ D = 1)\}$.
Applying Procedure 1 to find an FVTS, we obtain:
In Step 2, {16, 17} {52,53}→S.
In Step 3, {20, 21}, {48,49}→S
In Step 4, {24, 26}, {60,62}→S
In Step 5, {0, 1}, {8, 9}, {32,33}, {40,41}, {4, 5}, {12,13}, {36,37}, {44,45}→S
In Step 6, {19}, {55}→S.
In Step 7, none of the elements in S can be deleted.
The final combinations in S are starred in the map below. This is an FVTS of length 30.

0*	1*	2	3	4*	5*	6	7
8*	9*	10	11	12*	13*	14	15
16*	17*	18	19*	20*	21*	22	23
24*	25	26*	27	28	29	30	31
32*	33*	34	35	36*	37*	38	39
40*	41*	42	43	44*	45*	46	47
48*	49*	50	51	52*	53*	54	55*
56	57	58	59	60*	61	62*	63

VI GENERALIZATION AND AN UPPER BOUND ON THE LENGTH OF THE MINIMAL FVTS

Theorem 1 can be generalized so that a wider class of circuits can be treated. An upper bound on the length of the minimal FVTS will then be derived.

In Theorem 1 each basic element is considered separately to find the basic-element partitions. When k basic elements are considered at a time, a k^{th} order basic-element partition can be constructed.

Definition 8 Let J be the index set of k basic elements. The k^{th}order basic-element partition for $\{E_j: j \epsilon J\}$ is $P_J = [X|Y\|Z]$, where $Z = \bigcup_{j\epsilon J} IN(E_j) \cap \bigcup_{j\epsilon J} OUT(E_j)$, $X = \bigcup_{j\epsilon J} IN(E_j)-Z$, and $Y = \bigcup_{j\epsilon J} OUT(E_j)-Z$.

If k = 1, the first order partitions of Definition 3 are obtained.

Theorem 4 (based on Karp [12]).

Let $[X|Y\|Z]$ be a k^{th} order basic-element partition.

(a) For the complete set of comparable configurations,
$\{f_C(X,y_i,z): y_i \epsilon\ 2^{[Y]}\} \subseteq \{\psi_j(\alpha_1(X),\ldots,\alpha_k(X)): j = 1,2,\ldots,2^{2^k}\}$.

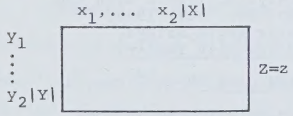
(b) For any $z \epsilon 2^{[Z]}$, the set of subfunctions $\{f_C(x_i;Y;z): x_i \epsilon 2^{[X]}\}$ contains at most 2^k distinct functions.

Proof: The partition defines a multiple nondisjoint decomposition of f_C. $f_C(X,Y,Z) = g(\beta_1(X,Z),\ldots,\beta_k(X,Z),Y,Z)$ where the β_i are the outputs of the k basic elements defining the partition.

(a) For $z \epsilon 2^{[Z]}$,
$\{f_C(X,y_i,z): y_i \epsilon 2^{[Y]}\}=\{g(\beta_1(X,z),\ldots,\beta_k(X,z),y_i,z)\}=\{\psi_i(\alpha_1(X),\ldots,\alpha_k(X))\}$
where $\alpha_i = \beta_i(X,z)$. But there are at most 2^{2^k} distinct functions of $\alpha_1,\ldots,\alpha_k$.

(b) For $z \epsilon 2^{[Z]}$,
$\{f_C(x_i,Y,z): x_i \epsilon 2^{[X]}\}=\{g(\beta_1(x_i,z),\ldots,\beta_k(x_i,z),Y,z): x_i \epsilon 2^{[X]}\}$,
where $(\beta_1(x_i,z),\ldots,\beta_k(x_i,z)) \epsilon \{0,1\}^k$. Thus, the set contains at most 2^k distinct functions, each of the form $g(u,Y,z)$ for $u \epsilon \{0,1\}^k$. This completes the proof.

Consider a complete set of comparable configurations $\{f_C(Xy_iz)\}$. These are the rows of the following map.



From Theorem 4(b), there are at most 2^k different columns. If there are exactly 2^k, then after verifying all function values appearing in one copy of each such column, we can verify all function values in any other column with at most 2^k-1 combinations. Thus, at most $2^k \cdot 2^{|Y|} + (2^k-1)\cdot(2^{|X|}-2^k)$ combinations are sufficient to verify $f_C(XYz)$. On the other hand, if there are fewer than 2^k distinct columns, consideration of the CFGs $\{f_C(Xy_iz): y_i \epsilon 2^{[Y]}\}$ in itself does not permit reduction in the number of combinations required; all $2^{[X\cup Y]}$ will be used.

From any k^{th} order basic-element partition $P_{kj}=[X|Y\|Z]$, an upper bound on the length of the minimal FVTS can be derived as follows.* If there are m maps $f_C(XYz)$, each with 2^k distinct columns, then the length of the minimal FVTS is no greater than

$$L_{kj} = m[2^k \cdot 2^{|Y|}+(2^k-1)\cdot(2^{|X|}-2^k)]+(2^{|Z|}-m)\cdot 2^{|X\cup Y|} \quad . \qquad (2)$$

Thus $\min_{1\le k\le t}\ \min_{\text{over all } P_{kj}} (L_{kj})$ is an upper bound on the length of the minimal FVTS of the circuit. In this formula, t is the largest integer solution to $t + 1 + \log_2(t + 1) \le |I|$, because for a map to have more than 2^k columns and 2^k distinct columns, k must satisfy $2^k < 2^{|X|}$, $k \le 2^{|Y|}$. The case $k = 2^{|Y|}$ leads to a map (subfunction) in which all combinations must be used since all k entries in each column must be known to classify it with respect to the 2^k distinct columns. Thus, if not all entries are to be used, k should satisfy $k + 1 \le 2^{|Y|}$. Therefore, it is necessary that $k + 1 + \log_2(k + 1) \le |X| + |Y| \le |I|$.

---

* The same bound can be obtained (with more difficulty) from consideration of rows instead of columns.

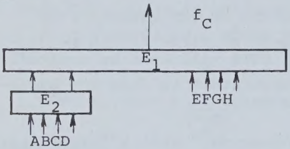
Let $[X|Y||Z]$ be a nontrivial first order basic-element partition ($k = 1$). Then an upper bound which is independent of the function realized on the structure is:

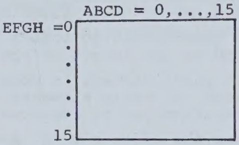
$$2\cdot 2^{|Y|} + (2^{|X|}-2) + (2^{|Z|}-1)\cdot 2^{|X\cup Y|} \qquad (3)$$

This follows from (2), for if the function is nondegenerate $m = 1$ is guaranteed because at least one $f_C(Xyz)$ is a ψ-CFG.

In the following example, we find an upper bound on an FVTS using a k^{th} order basic-element partition with $k = 2$.

Example 6 Consider the circuit





with the second order basic-element partition [ABCD|EFGH]. Assuming only that there exists 4 distinct columns in the map, an upper bound on the length of an FVTS is $2^2\cdot 2^4+(2^2-1)(2^4-2^2)=100$, whereas $2^{[I]} = 256$.

If the circuit had only a single line from E_2 to E_1, then the bound in (3) can be applied (assuming f depends nonvacuously on at least one x_i). At most 48 combinations are then required in an FVTS and, hence, in the most general multiple fault detection test. Note that only the structure of the circuit has been used to find this bound.

VII CONCLUSION

The approach introduced in this paper is expected to have application to a number of related problems, including the design of circuits which are easily tested, and new procedures for fault location. Further refinements to Procedure 1 may be possible by incorporating observations based on computational experience. It should also be possible to design an FVTS using k^{th} order basic-element partitions for $k > 1$. This should lead to even shorter tests and permit the treatment of circuits for which nontrivial partitions of first order do not exist.

REFERENCES

1. Schertz, D.R., "On the representation of digital faults," Ph.D. Thesis, Univ. of Ill., 1969. (AD 688-836)
2. Chang, H.Y., Manning, E. and Metze, G. Fault diagnosis of digital systems, John Wiley and Sons, New York, 1970.
3. Armstrong, D.B., "On finding a nearly minimal set of fault detection tests for combinational logic nets", IEEE Trans. on Elec. Computers, Vol. EC-15, 66-73, Feb. 1966.
4. Roth, J.P., "Diagnosis of automata failures: a calculus and a method," IBM J. of Research and Development, 278-291, July 1966.
5. Sellers, Jr., F.F., Hsiao, M.Y. and Bearnson, L.W., "Analyzing Errors with the Boolean Difference," IEEE Trans. on Computers, Vol. C-17, 676-683, July 1968.
6. Kautz, W.H., "Testing for faults in combinational cellular logic arrays," Conference Record, Eighth Annual Symposium on Switching and Automata Theory, 161-174, Oct. 1967.
7. Thurber, K.J., "Fault location in cellular arrays," AFIPS Conf. Proc., Vol. 35, 81-88, 1969.
8. Hornbuckle, G.D. and Spann, R.N., "Diagnosis of Single-Gate Failures in Combinational Circuits," IEEE Trans. on Computers, Vol. C-18, 216-220, March 1969.
9. Weiss, C.D., "Cell input-set systems and their use in fault detection for combinational networks: Part I: Fault detection tests for cell input-set systems", in review, January 1971.
10. Ashenhurst, R., "The decomposition of switching functions," Ann. Compt. Lab., Harvard Univ., Vol. 29, 30, 74-116, 1959.
11. Curtis, H.A., "A new approach to the design of switching circuits," D. Van Nostrand Company, Inc., Princeton, New Jersey, 1962.
12. Karp, R.M., "Functional decomposition and switching circuit design," SIAM J. Appl. Math., 11, 291-335, June 1963.
13. Roth, J.P., and Karp, R.M., "Minimization over Boolean graphs," IBM J. Research and Development, 4, 227-238, April 1962.

A METHOD FOR QUICK DETERMINATION OF THE PERIODIC STEADY-STATE IN NONLINEAR NETWORKS[1]

S. W. DIRECTOR, A. J. BRODERSEN and D. A. WAYNE
Department of Electrical Engineering
University of Florida
Gainesville, Florida 32601

I. ABSTRACT

Many nonlinear networks, are designed to operate in steady-state. If standard techniques are used to analyze this steady-state behavior, excessive computer time is wasted calculating the transient response. A Newton Raphson scheme based on the adjoint network method of calculating the jacobian can be used effectively to quickly determine the initial capacitance voltages and inductance currents which allows the network to start in steady state.

II. INTRODUCTION

The determination of periodic steady-state in electrical networks has always been of importance in the analysis of circuits since often this performance is critical for the evaluation of the circuit. For instance, the slew rate, a distortion which occurs at high frequencies and/or large voltage swings, of an operational amplifier can determine the useful band width of the amplifier. In computer simulations, the ability to quickly determine the periodic steady-state is necessary in the evaluation of the design. If the amplifier circuit is merely analyzed from time zero until the periodic steady-state is reached, many cycles may be required and the computational effort is likely to be excessive. This paper describes a computational algorithm which can efficiently determine steady-state.

Several methods have been used previously to determine steady state. Galerkin's method [1] is well known and has been exploited recently by Baily [2]. This method assumes that the form of the solution is a Fourier series of the fundamental frequency plus other harmonics. The coefficients of the Fourier series are then adjusted such that the differential equation is satisfied with the least error. More recently, Aprille and Trick [3] have aptly pointed out the shortcomings of this method: (1) Unless sufficient harmonics are included, the assumed solution may have little relation to the actual solution; (2) A mean square error function is used in determining the coefficient and the optimization techniques associated with the minimization of the error often have serious convergence problems;

---

[1] This research was supported in part by the National Science Foundation under Grant GK-27615.

(3) If a large number of harmonics are required, the number of variables to be optimized becomes excessive.

Apprille and Trick propose a method which uses a Newton iteration to determine the steady-state. Unfortunately, the algorithm is based upon a state space formulation of the network equations thus limiting the usefulness of their result. This paper describes a similar technique which is not dependent on the particular analysis scheme used. The method essentially adjusts the initial capacitance voltages and inductor currents until steady-state is reached by using a Newton iteration. The requisite Jacobian matrix is found by using the adjoint network concept.

III. DERIVATION OF THE METHOD

In what follows, it is convenient to associate all initial conditions with dc sources. Hence, a capacitor with initial voltage v_{CO} can be represented by a capacitor with zero initial voltage in series with a dc voltage source of value v_{CO} Similarly, an inductor with initial current i_{LO} can be represented by an inductor with zero initial current in parallel with a dc current source of value i_{LO}. These representations are shown in Fig. 1 where $v_C(t_o) = i_L(t_o) = 0$. Observe that if the initial conditions of a network are chosen such that the network is operating in the periodic steady-state with period T then

$$v_C(T+t_o) = v_C(t_o) = 0$$

and

$$i_L(T+t_o) = i_L(t_o) = 0.$$

Define $\underset{\sim}{x}$ to be the vector of initial condition source values:

$$\underset{\sim}{x} = (\underset{\sim}{v}_{CO}, \underset{\sim}{i}_{LO})^T$$

where $\underset{\sim}{v}_{CO}$ is the vector of initial capacitance source values and $\underset{\sim}{i}_{LO}$ is the vector of initial inductor source values. Clearly, the values of $\underset{\sim}{v}_C\ (T+t_o)$ and $\underset{\sim}{i}_C(T+t_o)$; depend upon the values of $\underset{\sim}{x}$. If we define

$$\underset{\sim}{f} = (\underset{\sim}{v}_C(T+t_o), \underset{\sim}{i}_C(T+t_o))^T$$

where $\underset{\sim}{v}_C(t)$ is the vector of capacitance voltages and $\underset{\sim}{i}_L(t)$ the vector of inductance currents, then

$$\underset{\sim}{f} = \underset{\sim}{f}(\underset{\sim}{x}).$$

We seek the value of $\underset{\sim}{x}$, $\underset{\sim}{x}^*$, such that

$$\underset{\sim}{f}(\underset{\sim}{x}^*) = \underset{\sim}{0}.$$

The Newton-Raphson [4] method may be used to iteratively determine a solution to this equation: In particular, upon application of the Newton Raphson method, we have that

$$\underset{\sim}{x}^{k+1} = \underset{\sim}{x}^{k} + \Delta\underset{\sim}{x}^{k}$$

where $\Delta\underset{\sim}{x}^{k}$ is the solution of

$$\underset{\sim}{J}^{k}\Delta\underset{\sim}{x}^{k} = -\underset{\sim}{f}(\underset{\sim}{x}^{k}).$$

$\underset{\sim}{J}^{k}$ is the jacobian

$$\underset{\sim}{J}^{k} = \frac{\partial\underset{\sim}{f}}{\partial\underset{\sim}{x}}\Bigg|_{\underset{\sim}{x} = \underset{\sim}{x}^{k}}.$$

When $\Delta\underset{\sim}{x}^{k}$ becomes sufficiently small the iterative process terminates.

It remains to determine the value of the jacobian. The adjoint network scheme can be fruitfully employed to ascertain the desired partial derivatives.

IV. A METHOD FOR JACOBIAN EVALUATION

We are interested in determining the variations of $v_C(t_o+T)$ and $i_L(t_o+T)$ to variations of v_{CO} and i_{LO}. Given the network we wish to analyze, and its adjoint [5], we can write, using Tellegen's theorem

$$\int_{t_o}^{t_o+T} \sum_{B} [v_B(t)\ \phi_B(\tau) - i_B(t)\ \psi_B(\tau)]\ dt = 0 \qquad (1)$$

where v_B and i_B denote the B^{th} branch voltage and current, respectively, in the original network, ϕ_B and ψ_B are the B^{th} branch voltage and current, respectively, in the adjoint network, and $\tau = T + 2t_o - t$. Expression (1) remains valid for incremental changes in v_B and i_B due to variations in the original network [5]

$$\int_{t_o}^{t_o+T} \sum_{B}[\Delta v_B(t)\ \phi_B(\tau) - \Delta i_B(t)\ \psi_B(\tau)]\ dt = 0. \qquad (2)$$

If the adjoint network elements are chosen properly, and if changes in v_B and i_B are due solely to perturbations of the initial condition sources, then (2) becomes

$$\sum_{CO}\int_{t_o}^{t_o+T} \phi_{CO}(\tau)\ dt\ \Delta v_{CO} - \sum_{LO}\int_{t_o}^{t_o+T} \psi_{LO}(\tau)\ dt\ \Delta\ i_{LO}$$

$$= \sum_{C} [C\ \psi_C(t_o)\ \Delta v_C(t_o+T) - C\,\psi_C(t_o+T)\ \Delta v_C(t_o)] \qquad (3)$$

$$- \sum_{L} [L\ \phi_L(t_o)\ \Delta i_L(t_o+T) - L\ \phi_L(t_o+T)\ \Delta i_L(t_o)]\ .$$

where the summations are taken over the indicated element types. But

$$v_C(t_o) = i_L(t_o) = 0$$

because we accounted for the initial conditions with independent sources. Therefore

$$\Delta v_C(t_o) = \Delta i_L(t_o) = 0.$$

We are free to arbitrarily choose values for the sources of the adjoint which correspond to the initial condition sources of the original. In particular, if we choose the value of the k^{th} capacitance initial condition to be

$$\psi_{C_k}(t_o) = \psi_{CO_k} = 1/C_k$$

with all other initial conditions set to zero:

$$\psi_{C_\ell}(t_o) = \psi_{CO_\ell} = 0, \text{ for all } \ell \neq k,$$

and

$$\phi_{L_\ell}(t_o) = \phi_{LO_\ell} = 0, \text{ for all } \ell,$$

then (3) becomes

$$\sum_{CO} \int_{t_o}^{t_o+T} \phi_{CO}(\tau)dt \ \Delta v_{CO} - \sum_{LO} \int_{t_o}^{t_o+T} \psi_{LO}(\tau)dt \ \Delta i_{LO}$$

$$= \Delta v_{C_k}(t_o+T). \tag{4}$$

From the chain rule

$$\Delta v_{C_k}(t_o+T) = \sum_{CO} \frac{\partial v_{C_k}(t_o+T)}{\partial v_{CO}} \Delta v_{CO} - \sum_{LO} \frac{\partial v_{C_k}(t_o+T)}{\partial i_{LO}} \Delta i_{LO}. \tag{5}$$

Comparison of (4) and (5) yield expressions for the partial derivatives:

$$\frac{\partial v_{C_k}(t_o+T)}{\partial v_{CO_\ell}} = \int_{t_o}^{t_o+T} \phi_{CO_\ell}(\tau)d\tau$$

$$= C_\ell[\phi_{C_\ell}(t_o+T) - \phi_{C_\ell}(t_o)]$$

because the current flowing through the initial condition source ψ_{CO_ℓ} also flows through the capacitor C_ℓ. Similarly

$$\frac{\partial v_{C_k}(t_o+T)}{\partial i_{LO_\ell}} = -\int_{t_o}^{t_o+T} \psi_{LO_\ell}(\tau)d\tau$$

$$= - L_\ell[\phi_{L_\ell}(t_o+T) - \phi_{L_\ell}(t_o)]$$

because the voltage across the initial condition source i_{LO_ℓ} equals the voltage across the inductor L_ℓ.

To find the partial derivatives of the final inductor currents with respect to the initial conditions we can choose

$$\phi_{LO_k} = -1/L_k$$

with all other initial conditions in the adjoint set to zero:

$$\phi_{LO_\ell} = 0, \text{ for all } 1 \neq k,$$

and

$$\psi_{CO_\ell} = 0, \text{ for all } 1.$$

Then (3) becomes

$$\sum_{CO} \int_{t_o}^{t_o+T} \phi_{CO}(\tau)dt\ \Delta v_{CO} - \sum_{LO} \int_{t_o}^{t_o+T} \psi_{LO}(\tau)dt\ \Delta i_{LO} = \Delta i_{L_k}(t_o+T)$$

from which we ascertain that

$$\frac{\partial \phi_{L_k}(t_o+T)}{\partial v_{CO_\ell}} = C_\ell[\psi_{C_\ell}(t_o+T) - \psi_{C_\ell}(t_o)]$$

and

$$\frac{\partial \phi_{L_k}(t_o+T)}{\partial i_{LO_\ell}} = - L_\ell[\phi_{L_\ell}(t_o+T) - \phi_{L_\ell}(t_o)].$$

We see from the above that for m energy storage elements a total of m+1 analysis are required to ascertain the m^2 entries of the Jacobian: one analysis of the original network and m analyses of the adjoint network under various excitations. These analyses are taken over one period.

As seen by the example which follows, the Jacobian need not be revaluated during each iteration. Rather, considerable computer time can be saved, if the Jacobian is only recomputed when no further improvement can be achieved.

V. EXAMPLE

The savings in computer time that may be achieved with the use of the technique described is illustrated by the simple example of Fig. 2, which was also used by Aprille and Trick [3]. The results are summarized in Table 1. The error, F, equals the sum of the difference between each capacitor voltage at the beginning and end of a cycle divided by the capacitor value and the inductor current at the beginning and end of a cycle divided by the inductor value. The results from continuous analyses are shown in the first three columns. Seventy-five cycles were required to achieve steady state and the analysis took 120 seconds of IBM 360/65 using a node based analysis which incorporated the variable step size-variable order Gear integration scheme [6]. The results using the method described above are shown in the columns 4-6. Note that approximately the same error was achieved in 4 iterations after 66 seconds of CPU time. During this run it was observed that the jacobian matrix didn't vary greatly from one iteration to the next. Because of this fact, it was decided to re-evaluate the jacobian only after there was no further reduction in F. The results of this run are shown in columns 7-9. Observe that although more iterations are required to achieve the desired value of F, less computer time (29 sec.) is needed--the reason being that the jacobian, for this example, had to be evaluated only once.

VI. REFERENCES

1. W. J. Cunningham, Nonlinear Analysis, New York: McGraw-Hill, 1958.

2. E. M. Baily, "Steady-state Harmonic Analysis of Nonlinear Networks -- A Computer-oriented Approach," Proc. First Ann. Houston Conf. on Circuits, Systems, and Computers, pp. 390-399, 1969.

3. T. J. Aprille and T. N. Trick, "Steady State Analysis of Nonlinear Circuits that have a Periodic Response," University of Illinois Coordinated Sciences Laboratory, Report R515, June 1971.

4. A. Ralston, A First Course in Numerical Analysis, McGraw Hill Book Company, 1965.

5. S. W. Director and R. A. Rohrer, "The Generalized Adjoint Network and Network Sensitivities," IEEE Trans. on Circuit Theory, Vol. CT-16, No. 3, pp. 318-322, August 1969.

6. C. W. Gear, "Simultaneous Numerical Solution of Differential-Algebraic Equations," IEEE Trans. on Circuit Theory, Vol. CT-18, No. 1, pp. 89-95, Jan. 1971.

CONTINUOUS INTEGRATION			ADJOINT TECHNIQUE			MODIFIED ADJOINT TECHNIQUE		
Cycle	F	CPU Time	Iterations	F	CPU Time	Iterations	F	CPU Time
1	7.88	3.00	1	1.97	17.35	1	1.97	17.05
2	3.41	5.05	2	0.4921	34.49	3	0.37	20.08
6	1.88	10.5	3	0.1782	51.08	7	3.88×10^{-2}	25.82
12	1.23	22.05	4	9.62×10^{-3}	66.2	10	7.17×10^{-3}	29.1
24	0.25	39.03						
50	0.160	80.86						
75	1.94×10^{-2}	120.1						

TABLE 1

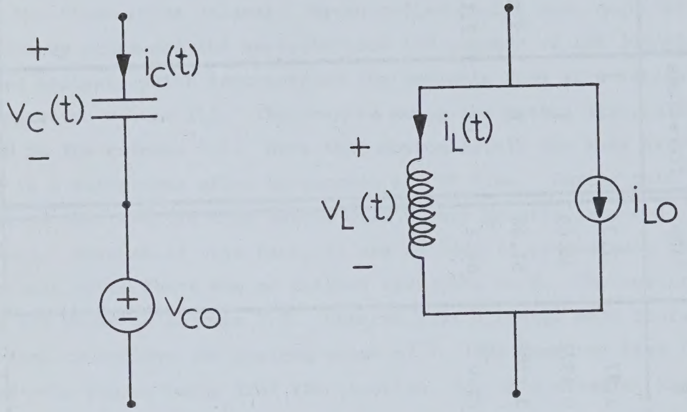
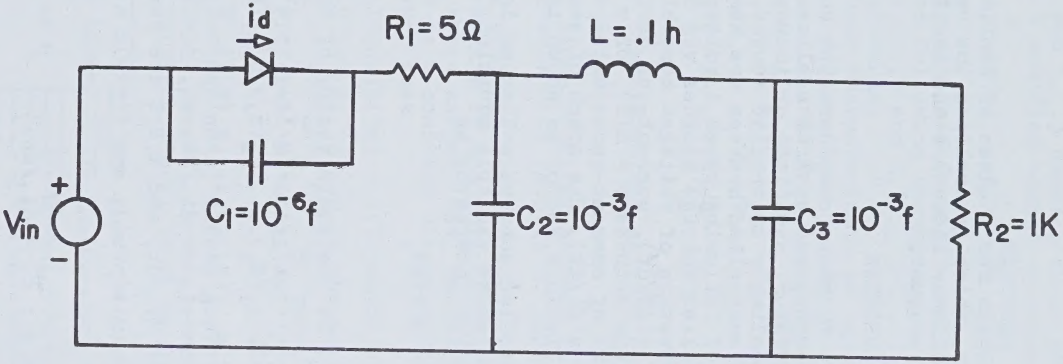

+
$i_C(t)$
$v_C(t)$
−
v_{CO}
$i_L(t)$
+
$v_L(t)$
−
i_{LO}

Fig. 1


id
R1=5Ω
L=.1h
C1=10^-6f
Vin
+
−
C2=10^-3f
C3=10^-3f
R2=1K

$$V_{in}=10\sin(60\cdot 2\pi\cdot t)$$

$$i_d=10^{-6}(\exp(40V_d)-1)$$

Fig. 2

ABSOLUTE STABILITY OF NETWORKS CONTAINING MULTIPLE NONLINEAR ACTIVE ELEMENTS

CHEN-YAO HO, AND YIH-GUANG JAN.
Dept. of Elecrical Engineering, Rice University,
Houston, Texas 77001.

ABSTRACT

Absolute stability conditions for a class of nonlinear networks are derived in this paper, which generalizes the results of [1] to the case where nonlinear inductors and nonlinear capacitors can be taken into account.

RESULTS

The class of network η under consideration consists of linear time-invariant elements and several nonlinear controlled sources (a nonlinear resistor, capacitor or inductor is a special case of such a nonlinear controlled source.) as shown in Fig. 1. The nonlinear controlled sources are memoryless, time-invariant, and are of following types : voltage- or derivative of voltage- controlled voltage sources $v = h(\tilde{v})$, or $v = h(d\tilde{v}/dt)$, voltage or derivative of voltage- controlled current sources $i = g(\tilde{v})$, or $i = g(d\tilde{v}/dt)$, current- or derivative of current- controlled voltage sources $v = h(\tilde{i})$, or $v = h(d\tilde{i}/dt)$, and current- or derivative of current- controlled current sources $i = g(\tilde{i})$, or $i = g(d\tilde{i}/dt)$. The graph of each of these functions is assumed to lie in a sector as shown in Fig. 2.

With nonlinear controlled sources acting as independent sources, we may write the state variable equations for η in the form we are particularly interested as

$$dx/dt = Ax + B\, f(z), \tag{1}$$

$$y = Cx. \tag{2}$$

where $x = \text{col.}(x_1,\dots,x_n)$ is the state vector of the network, $y = \text{col.}(\tilde{i}_1,\dots,\tilde{i}_\ell,\tilde{i}_{m+1},\dots,\tilde{i}_k,\tilde{v}_1,\dots,\tilde{v}_m,\tilde{v}_{\ell+1},\dots,\tilde{v}_p)$ is the output vector, $f(z) = \text{col.}(g_1(\bar{z}_1),\dots,g_\ell(\bar{z}_\ell),h_{m+1}(\bar{z}_{m+1}),\dots, h_k(\bar{z}_k),h_1(z_1),\dots,h_m(z_m),g_{\ell+1}(z_{\ell+1}),\dots,g_p(z_p))$, $\bar{z}_j$, $j = 1,\dots,\ell$ and $m+1,\dots,k$ is either $\tilde{i}_j$ or $d\tilde{i}_j/dt$, z_j, $j = 1,\dots,m$ and $\ell+1,\dots,p$ is either $\tilde{v}_j$ or $d\tilde{v}_j/dt$, and A,B,C are real constant and respectively $(n \times n)$, $(n \times (p+k))$, and $((p+k) \times n)$ matricas.

Define the sets

$$P_1 = \{\text{index } j : \bar{z}_j = d\tilde{i}_j/dt.\}$$

$$P_2 = \{\text{index } j : z_j = d\tilde{v}_j/dt.\}$$

and $P = P_1 \cup P_2$.

Then for every $j \in P$, we take the corresponding derivative of output variable in y. A typical one is

$$dy_j/dt = \sum_{i=1}^{n} c_{ij} \sum_{k=1}^{n} a_{ik} x_k + \sum_{i=1}^{n} c_{ij} \langle b_i, f(z) \rangle. \qquad (3)$$

where c_{ij}, a_{ik} are elements of C,A respectively and b_i is the i th row of B, and $\langle .,. \rangle$ denotes inner product. Following this way, we obtain a modified output equation for η as

$$dx/dt = Ax + Bf(z), \text{ and } z = \overline{C}x + \overline{D}f(z). \qquad (4)$$

where $\overline{C}, \overline{D}$ can be obtained from (2) and (3).

We make the following assumptions :

(A1) $g_j(.)$, $j=1,\ldots,p$ and $h_j(.)$, $j=1,\ldots,k$ are continuously differentiable.

(A2) The nonlinear characteristic of each controlled sources is confined to a sector, that is, there are positive constants $\alpha_1,\ldots,\alpha_\ell,\alpha_{\ell+1},\ldots,\alpha_k$ and $\beta_1,\ldots,\beta_m,\beta_{m+1},\ldots,\beta_p$ such that

$$0 \leq g_j(\tilde{i}_j)/\tilde{i}_j \text{ or } g_j(\dot{\tilde{i}}_j)/\dot{\tilde{i}}_j \leq \alpha_j, \quad j=1,\ldots,\ell$$
$$0 \leq g_j(\tilde{v}_j)/\tilde{v}_j \text{ or } g_j(\dot{\tilde{v}}_j)/\dot{\tilde{v}}_j \leq \alpha_j, \quad j=\ell+1,\ldots,k$$
$$0 \leq h_j(\tilde{v}_j)/\tilde{v}_j \text{ or } h_j(\dot{\tilde{v}}_j)/\dot{\tilde{v}}_j \leq \beta_j, \quad j=1,\ldots,m$$
$$0 \leq h_j(\tilde{i}_j)/\tilde{i}_j \text{ or } h_j(\dot{\tilde{i}}_j)/\dot{\tilde{i}}_j \leq \beta_j, \quad j=m+1,\ldots,p$$

(A3) The linear time-invariant network obtained by setting v_j, $j=1,\ldots,k$ and i_j, $j=1,\ldots,p$ to zero is asymptotically stable.

(A4) The linear time-invariant network is completely controllable and observable with respect to input and output ports.

By proving that $\lim_{t\to\infty} z(t) \to 0$ implies $\lim_{t\to\infty} y(t) \to 0$, we summarize the results in the form of the following

Theorem: Under above assumptions, the network η is absolutely stable if there is a constant diagonal matrix Q with nonnegative elements such that

$$\inf_{0 \leq w < \infty} \lambda_{\min} \left\{ \frac{1}{2} \left[(I+jwQ)G(jw) + ((I+jwQ)G(jw))^{*T} \right] + K^{-1} \right\} = \delta > 0$$

where $G(jw) = -\overline{C}(jwI-A)^{-1}B - \overline{D}$ is the modified transfer function of η and I is identity matrix, $\lambda_{\min.}$ denotes the lowest eigenvalues of $\{\ldots\}$ and $K = \text{diag.}(\alpha_1,..,\alpha_\ell,\beta_{m+1},..,\beta_p,\beta_1,..,\beta_m,\alpha_{\ell+1},\ldots,\alpha_k)$.

[1].C.Ho and R.deFigueiredo,"Global Asymptotical Stability of Networks Containing Several Nonlinear Controlled Sources," Proceeding of SWIEEECO, April 1970, Dallas Texas.

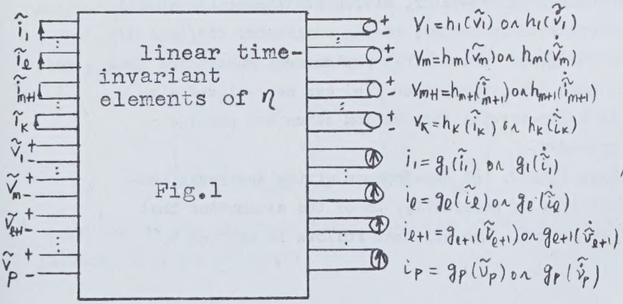

linear time-invariant elements of η
$\tilde{i}_1$
$\tilde{i}_\ell$
$\tilde{i}_{m+1}$
$\tilde{i}_k$
$\tilde{v}_1$
$\tilde{v}_m$
$\tilde{v}_{\ell+1}$
$\tilde{v}_p$
$v_1 = h_1(\tilde{v}_1)$ or $h_1(\dot{\tilde{v}}_1)$
$v_m = h_m(\tilde{v}_m)$ or $h_m(\dot{\tilde{v}}_m)$
$v_{m+1} = h_{m+1}(\tilde{i}_{m+1})$ or $h_{m+1}(\dot{\tilde{i}}_{m+1})$
$v_k = h_k(\tilde{i}_k)$ or $h_k(\dot{\tilde{i}}_k)$
$i_1 = g_1(\tilde{i}_1)$ or $g_1(\dot{\tilde{i}}_1)$
$i_\ell = g_\ell(\tilde{i}_\ell)$ or $g_\ell(\dot{\tilde{i}}_\ell)$
$i_{\ell+1} = g_{\ell+1}(\tilde{v}_{\ell+1})$ or $g_{\ell+1}(\dot{\tilde{v}}_{\ell+1})$
$i_p = g_p(\tilde{v}_p)$ or $g_p(\dot{\tilde{v}}_p)$

Fig.1

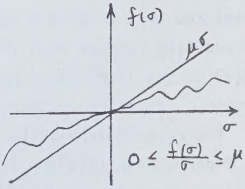


Fig.2.

MULTIPORT VOLTAGE GAIN FUNCTIONS: SYNTHESIS WITH ACTIVE RC MULTIPORT NETWORKS

THEODORE A. BICKART
Dept. of Electrical & Computer Engineering
Syracuse University
Syracuse, New York 13210

DONALD W. MELVIN
Dept. of Electrical Engineering
University of New Hampshire
Durham, New Hampshire 03824

ABSTRACT

A synthesis procedure is presented whereby any $q \times p$ matrix $\underset{\sim}{T}(s)$ of real rational functions of the complex frequency variable s can be realized as the voltage gain matrix of a (q+p)-port active RC network. The realization requires a minimum number of capacitors — $n = \text{degree } \{\underset{\sim}{T}(s)\}$ — and no more than $2(p + n)$ inverting, grounded voltage amplifiers or $p + n$ differential output, grounded voltage amplifiers. All capacitors and ports are grounded.

INTRODUCTION

Let N be a multiport network excited at p of its ports by voltages which are elements of the p-vector $\underset{\sim}{e}_s(t)$. Let the responses, the voltages at q of the ports, be elements of the q-vector $\underset{\sim}{e}_r(t)$. Let $\underset{\sim}{T}(s)$ be a $q \times p$ matrix of real rational functions of the complex frequency variable s such that

$$\underset{\sim}{E}_r(s) = \underset{\sim}{T}(s)\underset{\sim}{E}_s(s),$$

where $\underset{\sim}{E}_s(s) = \mathcal{L}\{\underset{\sim}{e}_s(t)\}$ and $\underset{\sim}{E}_r(s) = \mathcal{L}\{\underset{\sim}{e}_r(t)\}$.

A synthesis procedure is developed by which to realize $\underset{\sim}{T}(s)$ as an active RC multiport network with a minimum number of capacitors — $n = \text{degree } \{\underset{\sim}{T}(s)\}$ — and with no more than $2(p + n)$ inverting, grounded voltage amplifiers or $p + n$ differential output, grounded voltage amplifiers. Furthermore, the capacitors and ports will be grounded. The procedure is an extension of that developed by Melvin and Bickart [1] for realization of a short-circuit admittance matrix. Other extensions are described in [2]. References to the literature on active RC network synthesis of network functions will be found in both citations.

SYNTHESIS PROCEDURE

Suppose $\underset{\sim}{T}(s)$ is not regular at $s = \infty$ in the extended complex plane $\mathcal{C}^*$. Let $-\alpha \in \mathcal{C}^*$ be a point of regularity of $\underset{\sim}{T}(s)$ on the negative real axis. By invoking the Möbius transformation $s = \alpha z/(\beta - z)$ with $\beta > 0$, $\underset{\sim}{T}(s)$ is transformed into $\bar{\underset{\sim}{T}}(z) = \underset{\sim}{T}[\alpha z/(\beta - z)]$, which is regular at $z = \infty$ in $\mathcal{C}^*$. The Möbius transformation is a one-to-one mapping of $\mathcal{C}^*$ onto itself. Therefore,it's invertable; the inverse transformation is $z = \beta s/(s + \alpha)$. Now the results that follow establish a synthesis procedure which will realize $\bar{\underset{\sim}{T}}(z)$ as an active RC multiport network. In that network each capacitor, having admittance cz, should be replaced by a capacitor and resistor in series, having admittance $c\beta s/(s + \alpha)$; the resulting network will realize $\underset{\sim}{T}(s)$. Note: This replacement guarantees that if $\bar{\underset{\sim}{T}}(z)$ is realized with n grounded capacitors, then $\underset{\sim}{T}(s)$ can be realized with n grounded capacitors. This is a non-trivial observation since one premise was: The capacitors will be grounded.

Since $\text{degree } \{\underset{\sim}{T}(s)\} = \text{degree } \{\bar{\underset{\sim}{T}}(z)\}$, the development of the synthesis procedure can now continue, without loss of generality, under the assumption that $\underset{\sim}{T}(s)$ is regular at $s = \infty$. If it is not, then all that follows is applied to $\bar{\underset{\sim}{T}}(z)$ rather than $\underset{\sim}{T}(s)$.

Irreducible state equations associated with $\mathbf{T}(s)$ are of the form

$$\dot{\mathbf{x}} = \mathcal{F}\,\mathbf{x} + \mathcal{G}\,\mathbf{e}_s$$

$$\mathbf{e}_r = \mathcal{H}\,\mathbf{x} + \mathcal{J}\,\mathbf{e}_s, \tag{1}$$

where $\mathbf{x}$ is an n-vector. The set of matrices $\{\mathcal{F}, \mathcal{G}, \mathcal{H}, \mathcal{J}\}$ is not unique. In fact, the family of all such sets of matrices consists of the sets $\{\mathbf{Q}^{-1}\mathcal{F}_o\mathbf{Q}, \mathbf{Q}^{-1}\mathcal{G}_o, \mathcal{H}_o\mathbf{Q}, \mathcal{J}\}$, where $\mathbf{Q}$ is an arbitrary $n \times n$ non-singular real constant matrix and $\{\mathcal{F}_o, \mathcal{G}_o, \mathcal{H}_o, \mathcal{J}\}$ is some set in the family. One such set can be obtained by the Ho-Kalman algorithm; therefore, it is assumed that a set of irreducible state equations associated with $\mathbf{T}(s)$ is known.

The network that realizes $\mathbf{T}(s)$ will be the interconnection of a capacitive sub-network N_C and a resistive sub-network N_R, as in the network block diagram shown in Figure 1. Now, N_C is an n-port grounded, capacitive sub-network and N_R is a $(p + q + n)$-port grounded, resistive sub-network. Denote the

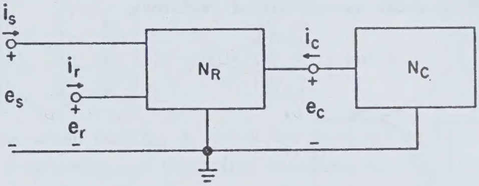


FIGURE 1 NETWORK BLOCK DIAGRAM

short-circuit parameter equations of N_R as

$$\begin{bmatrix} \mathbf{i}_s \\ \mathbf{i}_r \\ \mathbf{i}_c \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{ss} & \mathbf{G}_{sr} & \mathbf{G}_{sc} \\ \mathbf{G}_{rs} & \mathbf{G}_{rr} & \mathbf{G}_{rc} \\ \mathbf{G}_{cs} & \mathbf{G}_{cr} & \mathbf{G}_{cc} \end{bmatrix} \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \\ \mathbf{e}_c \end{bmatrix} = \mathbf{G} \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \\ \mathbf{e}_c \end{bmatrix}, \tag{2}$$

where $\mathbf{e}_c$ and $\mathbf{i}_c$ denote the n-vectors of voltages and currents at the ports common to N_R and N_C. The relation imposed on $\mathbf{e}_c$ and $\mathbf{i}_c$ will be expressed as

$$\mathbf{i}_c = -\,\mathcal{C}\,\dot{\mathbf{e}}_c. \tag{3}$$

Assume $\mathbf{G}_{rr}$ is non-singular. Then, in (2) with $\mathbf{i}_r = \mathbf{0}$, solve for $\mathbf{e}_r$ and $\mathbf{i}_c$ in terms of $\mathbf{e}_s$ and $\mathbf{e}_c$. Substitute (3) in that derived expression. The result, with $\mathbf{e}_c = \mathbf{x}$, yields (1) if and only if $\mathcal{C}$ is non-singular and

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{ss} & \mathbf{G}_{sr} & \mathbf{G}_{sc} \\ -\mathbf{G}_{rr}\mathcal{J} & \mathbf{G}_{rr} & -\mathbf{G}_{rr}\mathcal{H} \\ -\mathcal{C}\,\mathcal{G} - \mathbf{G}_{cr}\mathcal{J} & \mathbf{G}_{cr} & -\mathcal{C}\,\mathcal{F} - \mathbf{G}_{cr}\mathcal{H} \end{bmatrix}, \tag{4}$$

where, except for the fact that $\mathbf{G}_{rr}$ must be non-singular, the unspecified sub-matrices of $\mathbf{G}$ are arbitrary.

Thusfar, the n × n capacitance matrix $\underset{\sim}{C}$ is only constrained to be non-singular. Therefore, it can be assumed that $\underset{\sim}{C}$ is diagonal, with positive elements on the diagonal. By this assumption, the grounded n-port N_C contains n grounded capacitors — a characteristic of the realization of $\underset{\sim}{T}(s)$ that was to be established — with one capacitor connected to each port terminal, as illustrated by the schematic in Figure 2. This shows that not more than n capacitors are needed in the realization. That the realization requires at least n capacitors follows from the fact that $\underset{\sim}{C}$ would be singular if N_C contained fewer than n capacitors.

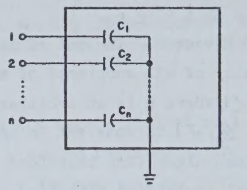


FIGURE 2 CAPACITIVE SUB-NETWORK N_C

The sub-network N_R in Figure 1, consisting of resistors and inverting, grounded voltage amplifiers, is assumed to have the structure shown in Figure 3, where $\bar{N}_R$ is a 3(p+q+n)-port grounded sub-network of resistors.

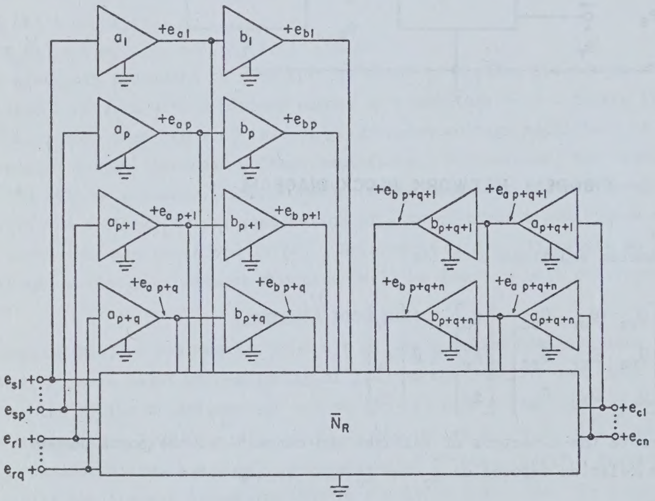


FIGURE 3 BLOCK DIAGRAM OF N_R

Denote the short-circuit parameter equations of $\bar{N}_R$ as

$$\begin{bmatrix} \underset{\sim}{i}_s \\ \underset{\sim}{i}_r \\ \underset{\sim}{i}_c \\ \underset{\sim}{i}_a \\ \underset{\sim}{i}_b \end{bmatrix} = \begin{bmatrix} \bar{\underset{\sim}{G}}_{ss} & \bar{\underset{\sim}{G}}_{sr} & \bar{\underset{\sim}{G}}_{sc} & \bar{\underset{\sim}{G}}_{sa} & \bar{\underset{\sim}{G}}_{sb} \\ \bar{\underset{\sim}{G}}'_{sr} & \bar{\underset{\sim}{G}}_{rr} & \bar{\underset{\sim}{G}}_{rc} & \bar{\underset{\sim}{G}}_{ra} & \bar{\underset{\sim}{G}}_{rb} \\ \bar{\underset{\sim}{G}}'_{sc} & \bar{\underset{\sim}{G}}'_{rc} & \bar{\underset{\sim}{G}}_{cc} & \bar{\underset{\sim}{G}}_{ca} & \bar{\underset{\sim}{G}}_{cb} \\ \bar{\underset{\sim}{G}}'_{sa} & \bar{\underset{\sim}{G}}'_{ra} & \bar{\underset{\sim}{G}}'_{ca} & \bar{\underset{\sim}{G}}_{aa} & \bar{\underset{\sim}{G}}_{ab} \\ \bar{\underset{\sim}{G}}'_{sb} & \bar{\underset{\sim}{G}}'_{rb} & \bar{\underset{\sim}{G}}'_{cb} & \bar{\underset{\sim}{G}}'_{ab} & \bar{\underset{\sim}{G}}_{bb} \end{bmatrix} \begin{bmatrix} \underset{\sim}{e}_s \\ \underset{\sim}{e}_r \\ \underset{\sim}{e}_c \\ \underset{\sim}{e}_a \\ \underset{\sim}{e}_b \end{bmatrix} = \bar{\underset{\sim}{G}} \begin{bmatrix} \underset{\sim}{e}_s \\ \underset{\sim}{e}_r \\ \underset{\sim}{e}_c \\ \underset{\sim}{e}_a \\ \underset{\sim}{e}_b \end{bmatrix} \qquad (5)$$

where the diagonal sub-matrices of $\bar{\underset{\sim}{G}}$ are symmetric and where $[\underset{\sim}{e}'_a \ \underset{\sim}{e}'_b]'$ and

$[\underset{\sim}{i}_a' \; \underset{\sim}{i}_b']'$ denote 2(p+q+n)-vectors of voltages and currents at the ports to which the outputs of the amplifiers are connected. Now

$$\begin{bmatrix} \underset{\sim}{e}_a \\ \underset{\sim}{e}_b \end{bmatrix} = \begin{bmatrix} \underset{\sim}{A} \\ \underset{\sim}{B}\,\underset{\sim}{A} \end{bmatrix} \begin{bmatrix} \underset{\sim}{e}_s \\ \underset{\sim}{e}_r \\ \underset{\sim}{e}_c \end{bmatrix} , \tag{6}$$

where $\underset{\sim}{A} = \text{diag}\{a_1, \ldots, a_{p+q+n}\}$ and $\underset{\sim}{B} = \text{diag}\{b_1, \ldots, b_{p+q+n}\}$. Note: Because the amplifiers are inverting amplifiers, each of the a_i and b_i is non-positive. Equations (5) and (6) together yield

$$\begin{bmatrix} \underset{\sim}{i}_s \\ \underset{\sim}{i}_r \\ \underset{\sim}{i}_c \end{bmatrix} = \left\{ \begin{bmatrix} \bar{\underset{\sim}{G}}_{ss} & \bar{\underset{\sim}{G}}_{sr} & \bar{\underset{\sim}{G}}_{sc} \\ \bar{\underset{\sim}{G}}_{sr}' & \bar{\underset{\sim}{G}}_{rr} & \bar{\underset{\sim}{G}}_{rc} \\ \bar{\underset{\sim}{G}}_{sc}' & \bar{\underset{\sim}{G}}_{rc}' & \bar{\underset{\sim}{G}}_{cc} \end{bmatrix} + \begin{bmatrix} \bar{\underset{\sim}{G}}_{sa} \\ \bar{\underset{\sim}{G}}_{ra} \\ \bar{\underset{\sim}{G}}_{ca} \end{bmatrix} \underset{\sim}{A} + \begin{bmatrix} \bar{\underset{\sim}{G}}_{sb} \\ \bar{\underset{\sim}{G}}_{rb} \\ \bar{\underset{\sim}{G}}_{cb} \end{bmatrix} \underset{\sim}{B}\,\underset{\sim}{A} \right\} \begin{bmatrix} \underset{\sim}{e}_s \\ \underset{\sim}{e}_r \\ \underset{\sim}{e}_c \end{bmatrix} . \tag{7}$$

By comparing this result with (2) it is found that

$$\underset{\sim}{G} = \bar{\underset{\sim}{G}}_p + \bar{\underset{\sim}{G}}_+ \underset{\sim}{A} + \bar{\underset{\sim}{G}}_- \underset{\sim}{B}\,\underset{\sim}{A} , \tag{8}$$

where

$$\bar{\underset{\sim}{G}}_p = \begin{bmatrix} \bar{\underset{\sim}{G}}_{ss} & \bar{\underset{\sim}{G}}_{sr} & \bar{\underset{\sim}{G}}_{sc} \\ \bar{\underset{\sim}{G}}_{sr}' & \bar{\underset{\sim}{G}}_{rr} & \bar{\underset{\sim}{G}}_{rc} \\ \bar{\underset{\sim}{G}}_{sc}' & \bar{\underset{\sim}{G}}_{rc}' & \bar{\underset{\sim}{G}}_{cc} \end{bmatrix} , \quad \bar{\underset{\sim}{G}}_+ = \begin{bmatrix} \bar{\underset{\sim}{G}}_{sa} \\ \bar{\underset{\sim}{G}}_{ra} \\ \bar{\underset{\sim}{G}}_{ca} \end{bmatrix} , \text{ and } \bar{\underset{\sim}{G}}_- = \begin{bmatrix} \bar{\underset{\sim}{G}}_{sb} \\ \bar{\underset{\sim}{G}}_{rb} \\ \bar{\underset{\sim}{G}}_{cb} \end{bmatrix} .$$

It will now be shown that $\bar{\underset{\sim}{G}}_p$, $\underset{\sim}{A}$, and $\underset{\sim}{B}$ can be determined such that $\bar{\underset{\sim}{G}}$ is hyperdominant — a necessary and sufficient condition for the 3(p+q+n)-port $\bar{N}_R$ to be realizable without internal nodes. Select $\bar{\underset{\sim}{G}}_p$ such that it is hyperdominant, with the sum of the elements in a row being strictly positive if the corresponding row of $\underset{\sim}{G} - \bar{\underset{\sim}{G}}_p$ has one or more non-zero elements. Let

$$\underset{\sim}{G} - \bar{\underset{\sim}{G}}_p = \underset{\sim}{P} + \underset{\sim}{M} , \tag{9}$$

where $\underset{\sim}{P}$ [$\underset{\sim}{M}$] contains the non-negative [non-positive] elements of $\underset{\sim}{G} - \bar{\underset{\sim}{G}}_p$. If $\bar{\underset{\sim}{G}}$ is to be hyperdominant, the elements of $\bar{\underset{\sim}{G}}_+$ and $\bar{\underset{\sim}{G}}_-$ must be non-positive. Lastly, the elements of $\underset{\sim}{A}$ and $\underset{\sim}{B}$ are non-positive. Therefore, (8) and (9) combine to give

$$\bar{\underset{\sim}{G}}_+ \underset{\sim}{A} = \underset{\sim}{P} \tag{10a}$$

and

$$\bar{\underset{\sim}{G}}_- \underset{\sim}{B}\,\underset{\sim}{A} = \underset{\sim}{M} . \tag{10b}$$

Clearly, by making the $|a_i|$ and the $|b_i|$ sufficiently large the non-zero elements of $\bar{\underset{\sim}{G}}_+$ and of $\bar{\underset{\sim}{G}}_-$ can be made as small in magnitude as is necessary to make the first p+q+n rows of $\bar{\underset{\sim}{G}}$ hyperdominant. This is possible because the sum of the elements in a row of $\bar{\underset{\sim}{G}}_p$ is strictly positive if the corresponding row of $\underset{\sim}{G} - \bar{\underset{\sim}{G}}_p$ has one or more non-zero elements. If $\underset{\sim}{A}^\dagger$ [$\underset{\sim}{B}^\dagger$] denotes the pseudo inverse of $\underset{\sim}{A}$ [$\underset{\sim}{B}$], then (10) yields

$$\bar{\underset{\sim}{G}}_+ = \underset{\sim}{P}\,\underset{\sim}{A}^\dagger , \tag{11a}$$

provided the consistency condition $\underset{\sim}{P}[\underset{\sim}{I} - \underset{\sim}{A}^\dagger \underset{\sim}{A}] = \underset{\sim}{0}$ is satisfied, and

$$\bar{\underset{\sim}{G}}_- = \underset{\sim}{M}\,\underset{\sim}{A}^\dagger \underset{\sim}{B}^\dagger , \tag{11b}$$

provided the consistency conditions $\underset{\sim}{M}[\underset{\sim}{I} - \underset{\sim}{A}^\dagger \underset{\sim}{A}] = \underset{\sim}{0}$ and $\underset{\sim}{M}\,\underset{\sim}{A}^\dagger[\underset{\sim}{I} - \underset{\sim}{B}^\dagger \underset{\sim}{B}] = \underset{\sim}{0}$ are satisfied. Note: The consistency conditions can be satisfied. One way of doing so is to select the a_i and the b_i to be strictly negative; then $\underset{\sim}{A}$ [$\underset{\sim}{B}$] will be non-singular and $\underset{\sim}{A}^\dagger = \underset{\sim}{A}^{-1}$ [$\underset{\sim}{B}^\dagger = \underset{\sim}{B}^{-1}$]. It then follows that

$[\underset{\sim}{I} - \underset{\sim}{A}^{\dagger}\underset{\sim}{A}] \equiv \underset{\sim}{0}$ and $[\underset{\sim}{I} - \underset{\sim}{B}^{\dagger}\underset{\sim}{B}] \equiv \underset{\sim}{0}$.

Now, let

$$\begin{bmatrix} \underset{\sim}{\bar{G}}_{aa} & \underset{\sim}{\bar{G}}_{ab} \\ \underset{\sim}{\bar{G}}'_{ab} & \underset{\sim}{\bar{G}}_{bb} \end{bmatrix}$$

be diagonal with the value of each diagonal element selected such that the sum of the elements in each of the last $2(p+q+n)$ rows of $\underset{\sim}{\bar{G}}$ is zero. This condition means that there will be no shunt resistors at the ports to which the outputs of the controlled sources are connected. This completes the specification of $\underset{\sim}{\bar{G}}$.

Thusfar, it has been established only that $2(p+q+n)$ inverting, grounded voltage amplifiers — the number of diagonal elements a_i and b_i of $\underset{\sim}{A}$ and $\underset{\sim}{B}$ — are sufficient. It will now be shown that $2(p+n)$ amplifiers are sufficient.

Consider (10). If j is the index of a zero column of $\underset{\sim}{M}$ [both $\underset{\sim}{M}$ and $\underset{\sim}{P}$], set $b_j = 0$ [both $a_j = 0$ and $b_j = 0$]. With these being the only a_i and b_i set equal to zero, the consistency conditions for equations (11) are satisfied. This is a consequence of the fact that $\underset{\sim}{A}^{\dagger} = \mathrm{diag}\{a_1^{\dagger}, \ldots, a_{p+q+n}^{\dagger}\}$, where $a_i^{\dagger} = 1/a_i$ if $a_i \neq 0$ and $a_i^{\dagger} = 0$ if $a_i = 0$, and that $\underset{\sim}{B}^{\dagger} = \mathrm{diag}\{b_1^{\dagger}, \ldots, b_{p+q+n}^{\dagger}\}$, where $b_i^{\dagger} = 1/b_i$ if $b_i \neq 0$ and $b_i^{\dagger} = 0$ if $b_i = 0$.

Now, set $\underset{\sim}{G}_{sr} = \underset{\sim}{0}$, $\underset{\sim}{G}_{sc} = \underset{\sim}{0}$, and $\underset{\sim}{G}_{cr} = \underset{\sim}{0}$. Let $\underset{\sim}{G}_{ss}$ be diagonal, let $\underset{\sim}{G}_{rr}$ be diagonal and non-singular, and let $\underset{\sim}{S}$ by symmetric, all such that $\underset{\sim}{G}_p = \mathrm{diag}\{\underset{\sim}{G}_{ss}, \underset{\sim}{G}_{rr}, \underset{\sim}{S}\}$ satisfies all previously imposed properties. It follows that columns $p+1$ through $p+q$ of both $\underset{\sim}{M}$ and $\underset{\sim}{P}$ will be zero. Hence, $b_i = 0$ and $a_i = 0$ for $i = p+1, \ldots, p+q$. As a consequence, no more than $2(p+n)$ inverting, grounded voltage amplifiers are needed.

AMPLIFIER GAIN SELECTION

The synthesis procedure was established on the observation that amplifier gains exist such that the equations in (10) have solutions corresponding to which the rows of $[\underset{\sim}{\bar{G}}_p\ \underset{\sim}{\bar{G}}_+\ \underset{\sim}{\bar{G}}_-]$ are hyperdominant. Now, some criteria will be given for choosing the amplifier gains.

Let $\mathcal{A}$ [$\mathcal{B}$] denote the set of indices of the non-zero a_i [b_i], and let $\mathcal{P}$ [$\mathcal{M}$] denote the set of indices of the rows of $\underset{\sim}{P}$ [$\underset{\sim}{M}$] having one or more non-zero elements. Let $\sigma_i \geq 0$ denote the sum of the elements in the i-th row of $\underset{\sim}{\bar{G}}_p$. Let $\pi_i \geq 0$ [$\mu_i \geq 0$] denote the sum of the elements in the i-th row of $\underset{\sim}{P}$ [$-\underset{\sim}{M}$]. Pick real constants ω_i such that $\omega_i = 0$ if $i \in \mathcal{M} - (\mathcal{P} \cap \mathcal{M})$, $0 < \omega_i < 1$ if $i \in \mathcal{P} \cap \mathcal{M}$, and $\omega_i = 1$ if $i \in \mathcal{P} - (\mathcal{P} \cap \mathcal{M})$.

Now, consider the case $a_i = a$ for $i \in \mathcal{A}$ and $b_i = b$ for $i \in \mathcal{B}$. It can be shown that if

$$-a \geq \max_{i\,\in \mathcal{P}} \{\pi_i/\sigma_i\omega_i\} \tag{12a}$$

and

$$-b \geq \frac{\max_{i\,\in \mathcal{M}} \{\mu_i/\sigma_i(1 - \omega_i)\}}{-a}, \tag{12b}$$

Then the rows of $[\underset{\sim}{\bar{G}}_p\ \underset{\sim}{\bar{G}}_+\ \underset{\sim}{\bar{G}}_-]$ will be hyperdominant.

In a particular situation it may be desirable to have all the amplifiers have the same gain. In that case $a_i = a$ for $i \in \mathcal{A}$ and $b_i = a$ for $i \in \mathcal{B}$. It follows from (12) that a should be chosen to satisfy the inequality

$$-a \geq \max\left\{[\max_{i \in \mathcal{P}} \{\pi_i/\sigma_i\omega_i\}],\ [\max_{i \in \mathcal{M}}\{\mu_i/\sigma_i(1 - \omega_i)\}]^{1/2}\right\}. \quad (13)$$

Suppose $b_i = -1$ for $i \in \mathcal{B}$. Then, for each $i \in \mathcal{B}$ the sub-network N_R contains a cascade of two amplifiers which, as illustrated in Figure 4, can be replaced by a single differential output amplifier. In addition, suppose

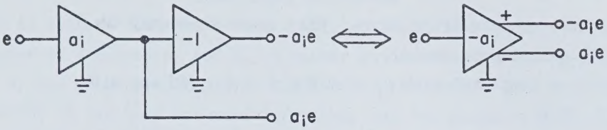


FIGURE 4 DIFFERENTIAL OUTPUT AMPLIFIER

$a_i = a$ for $i \in \mathcal{A}$. Then, it can be shown a must be selected such that

$$-a \geq \max_{i \in \mathcal{P} \cup \mathcal{M}}\{(\pi_i + \mu_i)/\sigma_i\}. \quad (14)$$

An immediate consequence of this way of choosing the amplifier gains is: No more than p+n differential output, grounded voltage amplifiers are needed.

EXAMPLE

The 1 × 1 matrix

$$\underset{\sim}{T}(s) = \left[\frac{1}{s^2 + \sqrt{2}\, s + 1}\right]$$

will be realized as a voltage gain matrix. Note: The network, once realized, will be a 2-pole Butterworth filter. The Ho-Kalman algorithm establishes the state equations which follow:

$$\dot{\underset{\sim}{x}} = \begin{bmatrix} 0 & 1 \\ -1 & -\sqrt{2} \end{bmatrix} \underset{\sim}{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \underset{\sim}{e}_s$$

$$\underset{\sim}{e}_r = \begin{bmatrix} 1 & 0 \end{bmatrix} \underset{\sim}{x} + \begin{bmatrix} 0 \end{bmatrix} \underset{\sim}{e}_s \ .$$

The state equations for equivalent networks will not be sought. Therefore, by (4), this set of matrices $\{\underset{\sim}{\mathcal{F}}, \underset{\sim}{\mathcal{G}}, \underset{\sim}{\mathcal{H}}, \underset{\sim}{\mathcal{J}}\}$ together with

$$\underset{\sim}{G}_{ss} = [1], \quad \underset{\sim}{G}_{sr} = [0], \quad \underset{\sim}{G}_{sc} = [0 \ \ -1], \quad \underset{\sim}{G}_{rr} = [1],$$

$$G_{cr} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \text{ and } \quad \underset{\sim}{\mathcal{C}} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}$$

yields

$$\underset{\sim}{G} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & -1/4 \\ -1 & 0 & 1 & \sqrt{2} \end{bmatrix} .$$

Note: $\underset{\sim}{G}_{sc}$ and $\underset{\sim}{G}_{cr}$ were set equal to zero matrices in demonstrating that at most 2(p+n) inverting amplifiers are needed; by making the choices indicated above, the actual number that will be required in this case is $2 < 2(p+n) = 6$.

Set

$$\bar{\underset{\sim}{G}}_p = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 2 & -1/4 \\ -1 & 0 & -1/4 & 2 \end{bmatrix} .$$

Then,

$$\underset{\sim}{P} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 5/4 & 0 \end{bmatrix} \quad \text{and } \underset{\sim}{M} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

have three — the same three — columns of zeros. This means that only two inverting amplifiers will actually be needed.

The realization will be completed with $b_i = -1$ for $i \in \mathcal{B} = \{3\}$ and with $a_i = a$ for $i \in \mathcal{A} = \{3\}$. Now,

i	σ_i	π_i	μ_i
1	0	0	0
2	0	0	0
3	3/4	0	1
4	$\sqrt{2} - 5/4$	5/4	0

Therefore, by (14),

$$-a \geq \max\{\tfrac{4}{3},\ 5/(4\sqrt{2} - 5)\} = 5/(4\sqrt{2} - 5) \approx 7.62.$$

With $a = -8$, (11) yields

$$\bar{\underset{\sim}{G}}_+ = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -5/32 & 0 \end{bmatrix} \quad \text{and } \bar{\underset{\sim}{G}}_- = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1/8 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} .$$

After composing $\bar{\underset{\sim}{G}}$ according to the guidelines previously developed, a schematic for $\underset{\sim}{T}(s)$ can be drawn; it is shown in Figure 5. A differential

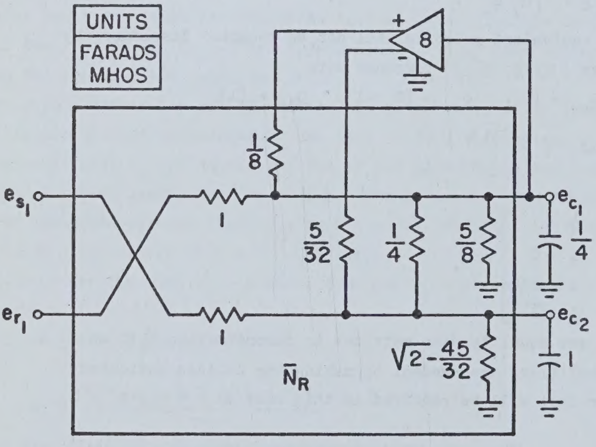

UNITS
FARADS
MHOS
+
-
8
$\frac{1}{8}$
e_{s_1}
e_{r_1}
1
1
$\frac{5}{32}$
$\frac{1}{4}$
$\frac{5}{8}$
e_{c_1}
$\frac{1}{4}$
e_{c_2}
1
$\sqrt{2}-\frac{45}{32}$
$\bar{N}_R$

FIGURE 5 EXAMPLE: REALIZATION OF $\underset{\sim}{T}(s)$

output amplifier is shown, rather than the cascade of the two amplifiers $a_3 = -8$ and $b_3 = -1$. Note: The 1 mho resistor in series with the response port terminal can be removed since the network realizes $\underset{\sim}{T}(s)$ with the response port open.

CONCLUDING DISCUSSION

It is evident that the actual number of amplifiers needed is equal to the number of non-zero a_i and b_i, a number which often can be reduced from $2(p+n)$ by the considerable design flexibility available. The diagonal elements of $\underset{\sim}{\mathcal{C}}$ can have any positive value, any non-singular Q can be used in $\{\underset{\sim}{\mathcal{F}}, \underset{\sim}{\mathcal{L}}, \underset{\sim}{\mathcal{H}}, \underset{\sim}{\mathcal{J}}\} = \{Q^{-1}\underset{\sim}{\mathcal{F}}_o\underset{\sim}{Q}, \underset{\sim}{Q}^{-1}\underset{\sim}{\mathcal{L}}_o, \underset{\sim}{\mathcal{H}}_o\underset{\sim}{Q}, \underset{\sim}{\mathcal{J}}\}$, and, except for $\underset{\sim}{G}_{rr}$, no preconditions exist on $\underset{\sim}{G}_{ss}$, $\underset{\sim}{G}_{sr}$, $\underset{\sim}{G}_{sc}$, $\underset{\sim}{G}_{rr}$, and $\underset{\sim}{G}_{cr}$.

The fact that the number of capacitors is a minimum and that all the capacitors and amplifiers are grounded is a distinct advantage if the network is to be fabricated as an integrated circuit.

The synthesis procedure was described for unloaded response ports — $\underset{\sim}{i}_r = \underset{\sim}{0}$. If the response ports are resistively loaded — $\underset{\sim}{i}_r = -\underset{\sim}{G}\,\underset{\sim}{e}_r$ — only a slight modification of the procedure is needed. In the line following (3), replace .... $\underset{\sim}{i}_r = \underset{\sim}{0}$ .... by .... $\underset{\sim}{i}_r = -\underset{\sim}{G}\,\underset{\sim}{e}_r$ ....

The synthesis procedure described herein can be reduced rather easily to a computer program which given $\underset{\sim}{T}(s)$ will generate $\underset{\sim}{\mathcal{C}}$, $\underset{\sim}{A}$, $\underset{\sim}{B}$, and $\bar{\underset{\sim}{G}}$.

REFERENCES

[1] D.W. Melvin and T.A. Bickart, "p-Port Active RC Networks: Short-Circuit Admittance Matrix Synthesis with a Minimum Number of Capacitors," IEEE Trans. on Circuit Theory, vol. CT-18, no. 6, November 1971.

[2] T.A. Bickart and D.W. Melvin, "Active RC Multiport Network Synthesis of Multiport Network Functions," To be published.

HIGHER ORDER LINEAR CIRCUIT MODELING

JOHN S. BEDFORD and HAROLD J. NEUMANN
The Pennsylvania State University, The Behrend Campus
Erie, Pennsylvania

ABSTRACT

A more accurate model of an electric circuit is possible by using components beyond the conventional capacitors, resistors, and inductors. The voltages across them are of the form $e_k = A_k d^k i/dt^k$ with k=0 resulting in a model of a resistor; k=1 an inductor; and k=-1 capacitor.

Higher order modeling, with $|k|>1$, allows for certain non-linearitaries which are frequency dependent, or in the mechanical analog, velocity dependent.

HISTORICAL AND PHILOSOPHICAL BACKGROUND

Historically, many laws of physics have been formulated by relating empirical data to mathematical relationships and the mathematical relationship is said to be a mathematical model of the physical system.

The relation e = L di/dt is an example. It appears to hold true. Another is that there is no divergence of flux density: $\nabla \cdot B = 0$. Occasionally, there is a report of someone trying to find an isolated magnetic pole. In this case it doesn't really matter if $\nabla \cdot B \neq 0$. If it doesn't, the conclusions reached by assuming it true are still valid for those systems in which it is zero. In fact, it would be interesting if someone would develop a theoretical electromagnetic system (perhaps, they have) where $\nabla \cdot B \neq 0$. The result might be that nothing is gained and that those who are looking for isolated poles, are looking for something of very little value. Of course it might go to the other way and there could develop an all out search to rival the searches for the quark, parity, etc.

The method of investigation here is: (1) postulate the existance of an entity; (2) determine if that entity is valuable; (3a) if valuable search for a physical realization or (3b) if not valuable, abandon the efforts. The method is not new but when it is applied, the results are usually reported only if valuable. (Who wants to report of their research that it was a waste of time?)

There have been a variety of proposed circuit elements and networks invented to simplify or increase the scope of circuit analysis. Some of which are nullators, norators[3], traditors[4], and unitors[5].

The objective here is to postulate the existance of higher order components other than capacitance, resistance, and inductance and examine the resulting circuit behavior. This general approach was applied here and higher order elements were investigated. At least one result was of value.

PROCEDURE

The three conventional components in a circuit are capacitance, C, resistance, R, and inductance, L. The voltages are respectively $(^1/_C)\int i dt$; Ri; and $L\, di/dt$.

To postulate the existence of higher order components it is sufficient to select a form of equation and proceed to analyze the resulting circuit behavior. The form selected here is convenient because of the symmetry of notation.

Assume additional higher order capacitance by writing

$$e_{C1} = (1/C1) \int i dt;\ e_{C2} = (1/C2)\int\int^{idtdt} \text{ etc. to}$$

$$e_{Ck} = (1/Ck)\int\cdot\cdot(k)\cdot\cdot\int idt\cdot\cdot\cdot(k)\cdot\cdot\cdot dt$$

and for higher order of L:

$$e_{L1} = L1\, di/dt;\ e_{L2} = L_2\, d^2i/dt^2;\text{ etc. to } e_{Lp} = L_p\, d^pi/dt^p$$

These can be written more compactly as $^en = N_n\ d^ni/dt^n$

where $d^{-q}i/dt^q$ is an alternate way to write $\int...(q)...\int i\ dt...(q)...dt$

Case example

The analysis of circuits containing higher order elements proceeds as usual. Consider, for example, the circuit of Fig. 1.

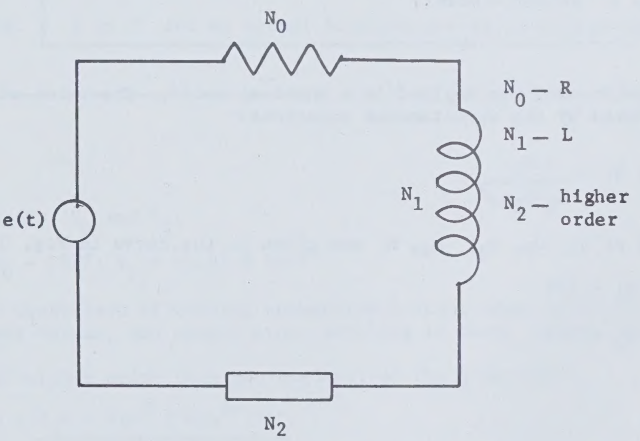


FIG. 1 Circuit for Second Order Inductor Model

$$N_0i + N_1\, di/dt + N_2\, d^2i/dt^2 = e(t)$$

The steady state solution for $e(t) = E \sin\omega t$ is $i(t) = (^{E}/_{Z}) \sin(\omega t - \phi)$

where $\phi = \tan^{-1} N_1\omega/(N_0 - N_2\omega^2)$

$Z = [(N_0 - N_2\omega^2)^2 + N_1{}^2\omega^2]^{1/2}$

The Q of this circuit is $N_1\omega/(N_0 - N_2\omega^2)$

The preceeding expression obtained for Q is very similar to that obtained partially imperically by Welsby[1] which is

$$Q = \frac{2\pi fL}{L(K + kf^2)}$$

Welsby's corresponding equivalent circuit with distributed capacitance neglected is shown in Fig. 2. He, in effect, separates the resistance into a D.C. part and A. C. part.

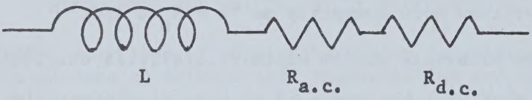


Fig. 2 Welsby's model

Comparison of Q

This expression of Q was applied to a physical coil[2]. The value of N_0 and N_2 were computed by the simultaneous equations:

$$Q_1 = \frac{\omega_1 N_1}{N_0 - N_2\omega_1{}^2} \; ; \; Q_2 = \frac{\omega_2 N_1}{N_0 - N_2\omega_2{}^2}$$

Where the values of Q_1, Q_2, ω_1, ω_2, $N_.$ are given in the curve in Fig. 3.

at f = 200 KHz $Q_1 = 750$

at f = 600 KHz $Q_2 = 700$

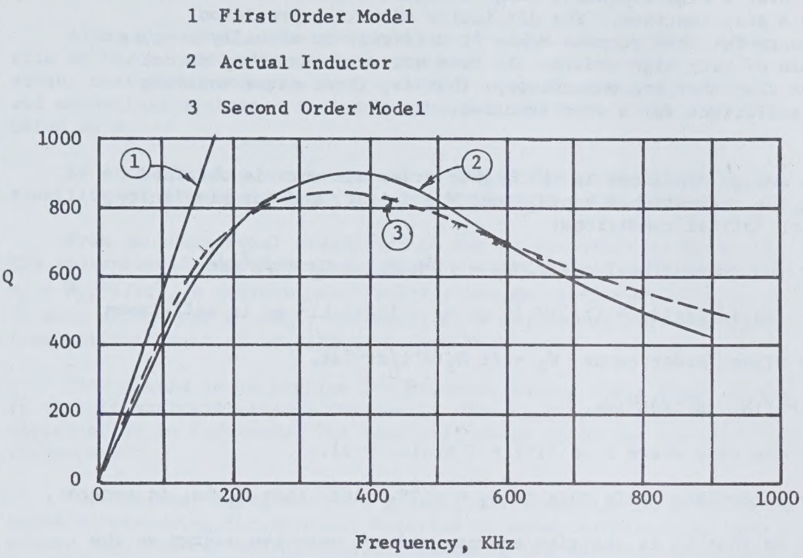


FIG. 3 Q vs f for an actual inductor and two different models

Solving for N_0 and N_2:

$N_0 = .572; \; N_2 = -1.41 \times 10^{-13}$

A comparison of Q using conventional first order modeling, measured values, and second order modeling is shown in Fig. 3.

If higher order than two are applied the Q becomes:

$$Q = \frac{N_1\omega - N_3\omega^3 + N_5\omega^5 - \cdots}{N_0 - N_2\omega^2 + N_4\omega^4 - \cdots}$$

Step Response

The analysis of response to a step function gives

$i(t) = (1/2A)(e^{b_1 t} - e^{b_2 t})$

$A = [(N_1/2N_2)^2 + (N_0/N_2)^2]^{1/2}$

$b_1 = -(N_1/2N_2)+A; \quad b_2 = -(N_1/2N_2)-A$

The value of b_1 is positive for second order modeling. This indicates that although second order modeling is sufficient for a steady sinusoid over a wide frequency range a higher order model must be used for a step function. The difficulty of making precision measurements for this purpose makes it difficult to actually compute the values of very high orders. It does not mean that they do not exist but that they are unnecessary; that is, first order modeling appears sufficient for a step function.

Energy

The energy contained in the higher order elements is obtained by using the conventional expression; $W = \int iedt$, and for simplicity, neglecting initial conditions:

For the conventional cases, $W_0 = \int i^2N_0dt$ or $W_R = \int i^2Rdt$

and $W_1 = \int iN_1(di/dt)dt = (1/2)N_1i^2$ or $W_L = (1/2)\ Li^2$ as is well-known.

For higher order cases, $W_2 = \int i\ N_2(d^2i/dt^2)dt$.

$W_n = \int iN_n\ (d^ni/dt^n)dt$.

For the case where i is $i(t) = I \sin(\omega t + \phi)$.

For a complete cycle this is $W_2 = -I^2N_2\pi$ note that N_2 is, in general, negative so that W_2 is positive energy, and not negative energy as the sign might indicate.

Similarly, $W_3 = \int -I^2\omega^3 \sin(\omega t + \phi) \cos(\omega t + \phi)dt$ and this is zero for a complete cycle.

If i(t) is a sinusoid the energy is zero for n odd and $(-1)^{(n/2)}\ N_nI^2\ \omega^{n-1}\pi$ for n even. This is to be expected as the energy for a complete cycle of a reactive component should be zero and not zero for a dissapative one.

Mechanical Application

Another application of this is in a mechanical system where the forces for a spring (k), mass (m), and friction (D) may be expressed as

$\int k\ vdt + Dv + m\ dv/dt = F$ applied.

By going one order higher the equation becomes

$\int kvdt + Dv + m\ dv/dt + Bd^2v/dt^2 = Fa$.

The Fourier transform at this yields an expression of the form

$D\omega - B\omega^3 + i(k + m\omega^2)$.

The friction, or dissapative part is $D\omega - B\omega^3$

Under some conditions the force, F, to sustain a velocity, v, is $F = av+bv^3$ where a and b are constants. The similarity of these two forms makes it reasonable to assume that higher order elements should simplify the analysis of some mechanical systems.

CONCLUSION

Although higher order elements may not exist in terms of being able to become merchantable items they can be said to exist to the extent that they produce a more accurate model for many electrical and mechanical systems. A closed conductor (shorted coil) approaches being an N_2.

By acknowledging the existence of the higher order elements the stability of some systems may be more accurately determined.

From an educational standpoint it may be desirable to appreciate the conventional components as part of a larger class. That is, with $e_k = N_k d^k i/dt^k$, a conventional inductor becomes a circuit element of $k = 1$. It does not become a really different type of element as capacitors are frequently thought of as different from inductors.

This should be no problem for students taking their first course in circuit analysis because frequently they have never seen either a capacitor or an inductor. The real difference is in the design of the components.

Furthermore, there may be a pedagogical advantage in thinking in terms of searching for physical entities to match mathematical models rather than the current process of searching for mathematical models to match or, nearly match, a physical process.

REFERENCES

1. Welsby, V. G. The Theory and Design of Inductance Coils,
   John Wiley & Sons 1960

2. Terman, F. E. Radio Engineers Handbook, third edition,
   McGraw-Hill 1943 (page 75)

3. Tellegen, B. D. H. On Nullators and Norators, IEEE Transactions on
   Circuit Theory V CT-13 n4 Dec. 66 p. 466-9

4. Duinker, S. Traditor, New Class of Non-energic Non-linear Network
   Elements Philips Research Reports V14 n1 Feb. 59 p.29-51

5. Keen, A. W. Proc. Inst. at Radio Engineers
   48, 248-250, 1960.

EXPERIMENTAL VERIFICATION OF SIGN REVERSAL OF TEMPERATURE COEFFICIENT

N.K. BOSE* & S.M. KOTOB
Department of Electrical Engineering
American University of Beirut
Beirut, Lebanon

ABSTRACT

An experimental set-up, involving a simulated lossless gyrator, a thermistor and a water-bath with thermometer, has been used to demonstrate the phenomenon of sign reversal of temperature coefficient, which is expected to have useful applications in temperature compensation circuiteries, among other possibilities.

A gyrator has already proved to be indispensable in various simulation problems[1]. Here, the impedance inverting property of a gyrator is used to show that sign reversal of temperature coefficients is possible in practice. It is well known that most metals exhibit an increase in resistance with increase in temperature, while two-terminal circuit elements like thermistors that are metallic oxide types of electronic semiconductors have relatively high negative temperature coefficients (n.t.c.) of resistance. Of course with the advanced solid-state technology at one's disposal, one can manufacture circuit elements, like a sensistor, by highly doping a semiconductor so that it behaves like a metal and has a positive temperature coefficient (p.t.c.) of resistance.[2] The present-day widespread use of these thermally sensitive resistors in applications ranging from vacuum manometers to oscillator frequency regulation and thermal infrared detection[3] motivate the experimental set-up described below.

A two-port, characterized by the open-circuit impedance matrix, $\begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix}$, and terminated at its second port by an impedance Z_L, has as driving-point impedance Z_1, at the first port,

$$Z_1 = z_{11} - \frac{z_{12}z_{21}}{z_{22} + Z_L} \tag{1}$$

For an ideal gyrator with $z_{11} = z_{22} = 0$, and gyration resistance

$r = z_{21} = -z_{12}$,

$$Z_1 = \frac{r^2}{Z_L} \tag{2}$$

which obviously implies that if Z_L decreases with temperature with r constant, then Z_1 must increase and vice-versa. The demonstration of this phenomenon necessitates simulation of a two-port with the open-circuit driving-point impedances at each port as close to zero as possible. A mismatch in the gyrator resistances (i.e. if $z_{12} = -r_1$, and $z_{21} = r_2$) will not alter the temperature reversal property, provided other factors remain unchanged. Among several gyrator circuits available, the Yanagisawa circuit with minor modifications was used to simulate a lossless gyrator. A KA31L1 thermistor by Fenwal was used as a load, and a water-bath with thermometer was used to vary and measure accurately the temperature.

---

* The first author is a visiting faculty member at the American University of Beirut from the University of Pittsburgh, Pittsburgh, Pa. 15213, U.S.A.

The temperature vs. resistance characteristic of the original thermistor, and of the device obtained by terminating the gyrator with it are shown plotted in the figure. The usual decaying exponential relationship of the thermistor resistance vs. its body temperature is shown to have been approximately changed into an exponentially increasing relationship. In essence, some sort of a sensistor has been simulated solely using electronic circuits. It is mentioned that if commercial simulation of sensistors from thermistors, or vice-versa, is desired, then the four bipolar transistors in the Yanagisawa gyrator circuit may be suitably replaced by using two f.e.t.'s. It must be borne in mind that in many applications of the simulated sensistor, a direct-coupled gyrator terminated by a thermistor is necessary. Another circuit by Yanagisawa[5] meets this demand.

CONCLUSIONS

In general, the feasibility of obtaining sign reversal of temperature coefficients can have important applications. In particular, the synthesis of a sensistor, using only electronic circuit components is interesting. The other possibility of sensistor fabrication using a highly doped semiconductor is usually not possible in ordinary laboratories.

REFERENCES

1. S.K. Mitra, "Synthesizing active filters," IEEE Spectrum vol. 6, Jan. 1969, pp. 47-63.

2. J. Millman and C.C. Halkias, "Electronic Devices and Circuits," McGraw-Hill Book Co., N.Y. 1967, pp. 287-288.

3. R.T. Chevako, "Thermistor bibliography," Proc. of IEEE, vol. 53, Feb. 1965, pp. 175-176.

4. T. Yanagisawa, "Realisation of a lossless transistor gyrator," Electronic Letters, vol. 3, April 1967, pp. 167-168.

5. T. Yanagisawa, "Direct-coupled transistor lossless gyrator," Electronic Letters, vol. 3, July 1967, pp. 311-312.

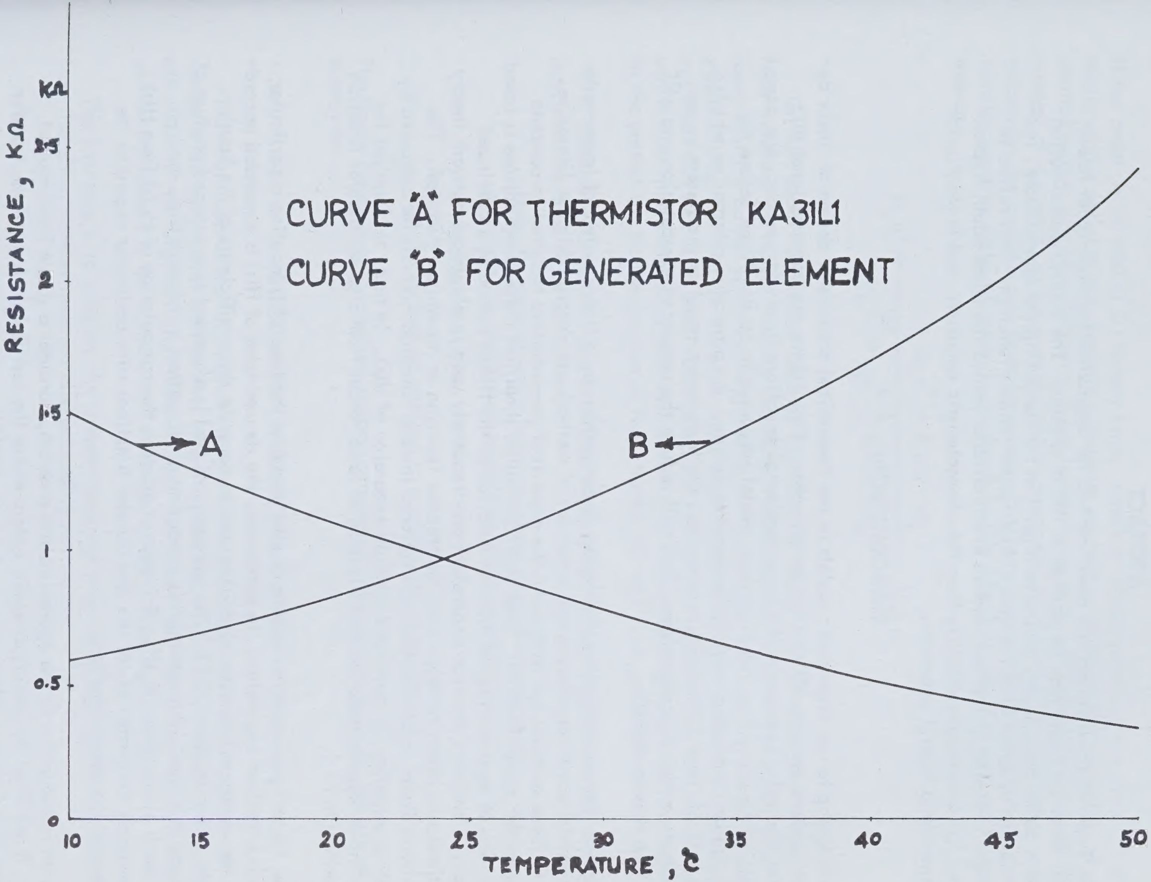


FIG. RESISTANCE - TEMPERATURE CHARACTERISTICS

ROUTH APPROXIMATION METHOD FOR HIGH-ORDER LINEAR SYSTEMS

MAURICE F. HUTTON
The Singer Company • Kearfott Division, Little Falls, New Jersey

ABSTRACT

The 'Routh Approximation' is a novel method for approximating high-order linear systems where the order may be as high as 100 or greater. The Routh Approximation method is based on a finite, recursive algorithm for computing the approximate, low-order transfer function from the given, high-order transfer function. This recursive algorithm is related to the well-known Routh criteria (hence the name Routh Approximation) for determining stability from the characteristic equation and is readily implemented on a digital computer.

INTRODUCTION

Attempting to use high-order models in the theoretical analysis or design of linear dynamical systems presents difficulties in two areas: First, there are computational difficulties that arise because of the large number of equations involved; second, the complex design resulting from the high-order model may have to be simplified before the design can be considered feasible for implementation. An attractive alternative which avoids both of these difficulties is to replace the high-order model of the system by an approximate model of lower-order. This will reduce the amount of computation and will result in a simpler design.

Methods for approximating high-order linear systems by a linear system of lower-order are certainly worth considering and a number of methods are described in the literature. Many of these methods are limited by the numerical computations required to compute the low-order approximation. One approach which simplifies these computations is based on the idea of representing the transfer function of the linear system as a continued fraction. Continued fraction expansions are frequently used in electrical network theory to synthesize a linear network from the transfer function or impedance function. The well-known Pade' approximation of a rational transfer function H(s) can be computed by truncating a particular continued fraction expansion of H(s). In fact, this method for computing an approximate transfer function of lower-order from a given transfer function is described in [1].

The Routh Approximation method is also based on the idea of truncating a particular continued fraction expansion. In particular, the denominator of H(s) is expanded according to the continued fraction algorithm used to compute the coefficients of the Routh stability criterion (See [2]). The numerator of H(s) is expanded as a sum and product of continued fractions. Truncating this expansion of H(s) after k terms defines the k-th order Routh approximant $H_k(s)$. This paper describes the computation of $H_k(s)$ from H(s) and compares the properties of the two transfer functions with particular regard to the accuracy of the approximation.

As an introduction to the approximation problem, consider a single input-output, linear, fixed (time invariant) dynamic system where the coefficient form of the transfer function H(s) describing the system as shown in (1) is known.

$$H(s) = \frac{b_1 s^{n-1} + \ldots + b_n}{a_0 s^n + a_1 s^{n-1} + \ldots + a_n} \qquad (1)$$

If the order n of H(s) in (1) is very large, possibly 100 or greater, then using (1) directly as the model of the dynamic system for analysis purposes or for the design of a control system may be a difficult, if not impractical, task. In many cases a considerable number of the modes in the partial fraction expansion of H(s) may be neglected. Consequently, for analysis and design purposes, H(s) may be replaced by an approximate transfer function $H_k(s)$ of order k, as shown in (2), where k is substantially smaller than n.

$$H_k(s) = \frac{d_1 s^{k-1} + \ldots + d_k}{c_0 s^k + c_1 s^{k-1} + \ldots + c_k} \qquad k \leq n \qquad (2)$$

Let h(t) equal the inverse Laplace transform of H(s) and $h_k(t)$ equal the inverse Laplace transform of $H_k(s)$. Stating that $H_k(s)$ "approximates" H(s) will be considered to mean that the output $y_k(t)$ resulting from $h_k(t)$ for an input u(t) tends to match the output y(t) resulting from h(t) for the same input u(t). In other words, the error $\epsilon(t)$ between $y_k(t)$ and y(t) should be small relative to y(t). It will be assumed that the class of inputs u(t) is such that if the poles and residues of $H_k(s)$ closely approximates the dominant modes in the partial fraction expansion of H(s), then $y_k(t)$ will be a good approximation of y(t).

In developing a general approximation method to compute $H_k(s)$ from H(s), the following questions need to be considered:

- Does the method result in an accurate approximation which preserves the significant properties of the original system?
- Is the method easy to use?
- Can the method be applied to a wide class of problems?

The Routh Approximation method appears promising with respect to all of the above questions.

ROUTH APPROXIMATION METHOD

The Routh approximation requires only a finite number of arithmetic operations to compute the polynomial coefficients of the approximant $H_k(s)$ from the polynomial coefficients of the original, high-order transfer function H(s). The Routh approximation can be considered to be composed of three basic steps, described below.

The first step is to compute the transfer function $\hat{H}(s)$ for a different linear system from the given transfer function H(s) according to

$$\hat{H}(s) = \frac{1}{s} H\left(\frac{1}{s}\right) \quad . \qquad (3)$$

The transformation from H(s) to $\hat{H}(s)$ is referred to as the reciprocal transformation since a nonzero pole p (or zero z) is mapped into the pole $1/p$ (or zero $1/z$) of $\hat{H}(s)$. Using

(3), the reciprocal transfer function corresponding to H(s) in (1) is

$$\hat{H}(s) = \frac{b_1 s^{n-1} + \ldots + b_1}{a_n s^n + \ldots + a_1 s + a_0} \tag{4}$$

It is evident from (4) that the reciprocal transformation is simply the operation of reversing the order of the polynomial coefficients in the transfer function.

The second step is to expand the transfer function $\hat{H}(s)$ in terms of n scalers $(\alpha_1, \alpha_2, \ldots \alpha_n)$, called the alpha parameters, and a corresponding set of n scalers $(\beta_1, \beta_2, \ldots \beta_n)$, called the beta parameters. Starting from the coefficient form of the transfer function $\hat{H}(s)$ shown in (4), the recursive expansion for the alpha and beta parameters is similar to a continued fraction expansion, in particular, the resulting form is the sum and product of continued fractions. To illustrate, the resulting expansion in the case of a 4th order transfer function is

$$\begin{aligned} \hat{H}(s) = \left(\frac{1}{1+\alpha_1 s} + \frac{1}{\alpha_2 s} + \frac{1}{\alpha_3 s} + \frac{1}{\alpha_4 s}\right) \Bigg[\beta_1 \\ + \left(\frac{1}{\alpha_2 s} + \frac{1}{\alpha_3 s} + \frac{1}{\alpha_4 s}\right) \Bigg[\beta_2 \\ + \left(\frac{1}{\alpha_3 s} + \frac{1}{\alpha_4 s}\right) \Bigg[\beta_3 \\ + \left(\frac{1}{\alpha_4 s}\right) \Big[\beta_4 \Big] \Bigg] \Bigg] \Bigg] \end{aligned} \tag{5}$$

where the parenthetical terms in (5), such as

$$\left(\frac{1}{1+\alpha_1 s} + \frac{1}{\alpha_2 s} + \frac{1}{\alpha_3 s} + \frac{1}{\alpha_4 s}\right)$$

are the symbolical form of a Stieltjes continued fraction expansion used in place of the usual form

$$\cfrac{1}{1+\alpha_1 s + \cfrac{1}{\alpha_2 s + \cfrac{1}{\alpha_3 s + \cfrac{1}{\alpha_4 s}}}}$$

As shown in [2], the above Stieltjes continued fraction expansion defines an algorithm for computing the coefficients $\alpha_1, \alpha_2, \ldots \alpha_n$ of the Routh stability criterion (system is asymptotically stable if and only if $\alpha_1 > 0, \alpha_2 > 0, \ldots, \alpha_n > 0$) from the denominator of the transfer function. It is clear from (5) what the general form of the alpha-beta expansion would be for an n-th order transfer function. An iterative procedure for computing the alpha-beta expansion from the coefficient form of the transfer function will be illustrated by an example subsequently. For the k-th order Routh approximation, the alpha-beta expansion of the reciprocal transfer function is terminated after k steps, and the truncated expansion is summed to give the k-th order transfer function $\hat{H}_k(s)$.

$$\hat{H}_k(s) = \frac{d_k s^{k-1} + \dots + d_1}{c_k s^k + \dots + c_1 s + c_0} \tag{6}$$

If, for example, $\hat{H}(s)$ is 4th order and its alpha-beta expansion is given by (5), then for a 2nd order approximation

$$\hat{H}_2(s) = \left(\frac{1}{1+\alpha_1 s} + \frac{1}{\alpha_2 s}\right)\left[\beta_1 + \left(\frac{1}{\alpha_2 s}\right)\left[\beta_2\right]\right] \tag{7}$$

which, when rearranged gives

$$\hat{H}_2(s) = \frac{d_2 s + d_1}{c_2 s^2 + c_1 s + c_0} \tag{8}$$

The third step is to compute the coefficient form of the transfer function $H_k(s)$ from $\hat{H}_k(s)$ in (6) by again applying the reciprocal transformation defined by (3). The resulting k-th order transfer function $H_k(s)$ shown in (2) is defined to be the k-th order Routh approximant. In the example of the 2nd order approximation described by (7) and (8), the Routh approximant is therefore

$$H_2(s) = \frac{d_1 s + d_2}{c_0 s^2 + c_1 s + c_2} \tag{9}$$

The three basic steps of the Routh approximation are graphically represented (solid arrows) in Figure 1.

In order to illustrate the details of the Routh approximation, consider the following numerical example: Suppose

$$H(s) = \frac{28 s^3 + 496 s^2 + 1800 s + 2400}{2 s^4 + 36 s^3 + 204 s^2 + 360 s + 240}, \tag{10}$$

then

$$\hat{H}(s) = \frac{2400 s^3 + 1800 s^2 + 496 s + 28}{240 s^4 + 360 s^3 + 204 s^2 + 36 s + 2}. \tag{11}$$

Next, $\hat{H}(s)$ is rewritten in the form

$$\hat{H}(s) = \frac{\dfrac{2400 s^3 + 1800 s^2 + 496 s + 28}{360 s^3 + 36 s}}{1 + \dfrac{240 s^4 + 204 s^2 + 2}{360 s^3 + 36 s}} \tag{12}$$

where the improper rational functions in the numerator and denominator of (12) are expanded to give

$$\hat{H}(s) = \frac{\dfrac{20}{3} + \dfrac{1800\, s^2 + 256\, s + 28}{360\, s^3 + 36\, s}}{1 + \dfrac{2}{3} + \dfrac{180\, s^2 + 2}{360\, s^3 + 36\, s}} \; . \tag{13}$$

In order to continue the expansion, (13) is rewritten in the form

$$\hat{H}(s) = \frac{1}{1 + \dfrac{2}{3}\, s + \dfrac{1}{\dfrac{360\, s^3 + 36\, s}{180\, s^2 + 2}}} \left[\frac{20}{3} + \frac{1}{\dfrac{360\, s^3 + 36\, s}{180\, s^2 + 2}} \left[\frac{1800\, s^2 + 256\, s + 28}{180\, s^2 + 2} \right] \right] . \tag{14}$$

From comparing (14) to (5), one immediately gets that

$$\alpha_1 = \frac{2}{3} \, , \qquad \beta_1 = \frac{20}{3} \; . \tag{15}$$

The improper rational functions in (15) are then expanded as before to give

$$\frac{360\, s^3 + 36\, s}{180\, s^2 + 2} = 2\, s + \frac{32\, s}{180\, s^2 + 2} \, ,$$

and

$$\frac{1800\, s^2 + 256\, s + 28}{180\, s^2 + 2} = 10 + \frac{256\, s + 8}{180\, s^2 + 2}$$

yielding

$$\alpha_2 = 2 \, , \qquad \beta_2 = 10 \; . \tag{16}$$

Continuing in this manner, the remaining alpha and beta parameters can be computed. Substitution of (15) and (16) into (7) results in

$$\hat{H}_2(s) = \frac{30\, s + 40}{3\, s^2 + 6\, s + 4} \; .$$

The 2nd order Routh approximant to (10) is, therefore,

$$H_2(s) = \frac{40\, s + 30}{4\, s^2 + 6\, s + 3} \; .$$

As an aid to the computations, a pair of tables, referred to as the denominator array and the numerator array, can be constructed to compute the alpha and beta parameters, respectively. For the transfer function $\hat{H}(s)$ given by (11), the pair of arrays are

Denominator Array

	240	204	2
	360	36	
$\alpha_1 = 2/3$	180	2	
$\alpha_2 = 2$	32		
$\alpha_3 = 45/8$	2		
$\alpha_4 = 16$			

Numerator Array

	2400	496
	1800	28
$\beta_1 = 20/3$	256	
$\beta_2 = 10$	8	
$\beta_3 = 8$		
$\beta_4 = 4$		

A detailed discussion of the algorithms for computing the denominator and numerator arrays to obtain the alpha and beta parameters, respectively, from the polynomial coefficients of the transfer function, is given in [3]. As one might suspect from the title of the method, 'Routh Approximation', the denominator array (that part to the right of the vertical line) is precisely the well-known Routh table for determining if the transfer function $\hat{H}(s)$ in (11) is stable.

PROPERTIES OF THE ROUTH APPROXIMANT

Several general properties of the Routh approximants, which in part explain the accuracy of the method, are discussed without proof* in this section.

Before discussing the general properties, it is instructive to first indicate the accuracy of this method on another simple 4th order linear system whose transfer function is

$$H(s) = \frac{14.7\,s^3 + 1363.38\,s^2 + 28566.8\,s + 92024.7}{s^4 + 97\,s^3 + 2268\,s^2 + 11680\,s + 14000} . \qquad (17)$$

The approximate transfer functions $H_k(s)$, $k = 1,2,3,4$ corresponding to the 1st, 2nd, 3rd, and 4th order Routh approximations, respectively, are shown in Table 1 in partial fraction form (obviously, $H_4(s) = H(s)$). An examination of the poles and residues in Table 1 clearly demonstrates the capability of the Routh approximation to accurately approximate the dominant modes of the partial fraction expansion.

Since the denominator array for computing the alpha parameters is the Routh table for determining stability from the characteristic equation, the alpha parameters have the property that a linear system is asymptotically stable (all eigenvalues have negative real parts) if and only if all the alpha parameters of that system (or the corresponding reciprocal system) are positive. The alpha parameters of the k-th order Routh approximant are by definition a subset of the alpha parameters of the given, high-order transfer function (in the reciprocal domain). Hence, it can be stated that if the given, high-order model of the system is asymptotically stable, then the approximate, low-order model computed by the Routh approximation will also be asymptotically stable. In other words, the Routh approximation preserves stability.

---

* Proofs of these properties will be included in subsequent publications.

Another property which indicates the accuracy of the Routh approximants results from comparing the power (or Talyor) series expansion of $H_k(s)$ in s about $s = 0$ to the corresponding series expansion of $H(s)$. It can be shown that the coefficients in the two power series are identical for at least the first k terms in the expansion. Consequently, the frequency response of $H_k(s)$ tends to match the frequency response of $H(s)$ for low frequencies.

Let $h(t)$ denote the unit impulse response of a linear system, then the impulse response "energy" is defined by the integral

$$I = \int_0^\infty h^2(t)\,dt \, .$$

A useful criteria for estimating the accuracy of the Routh approximation is to compare the impulse response energy of the approximate transfer function to the impulse response energy of the exact transfer function. This is easy to do as part of the Routh approximation as a consequence of a theorem derived by Astrom [3] and the fact that the impulse response energies for the original and reciprocal transfer functions are identical. The impulse response energy I_k corresponding to the k-th order Routh approximation of a given n-th order transfer function is

$$I_k = E_1 + E_2 + \dots + E_k \qquad k = 1, 2, \dots, n$$

where

$$E_i = \frac{\beta_i^2}{2\alpha_i} \tag{18}$$

and where

$$\{\alpha_1, \alpha_2, \dots, \alpha_n, \beta_1, \beta_2, \dots, \beta_n\}$$

is the sequence of alpha and beta parameters used in the computation of the Routh approximation. Since E_i in (18) is always non-negative for a stable system, the sequence of impulse response energies $I_1, I_2, \dots, I_n$ is non-decreasing and bounded from above by the impulse response energy I of the given n-th order transfer function, i.e.

$$I_1 \le I_2 \le \dots \le I_n = I \, .$$

Therefore, the ratio

$$\frac{I_k}{I} = \frac{E_1 + E_2 + \dots + E_k}{E_1 + E_2 + \dots + E_n} \qquad k = 1, 2, \dots, n \tag{19}$$

provides a useful technique for selecting the order k of the approximation to be used. For instance, the criteria might be to pick the smallest k for which the ratio in (19) is greater than 90 percent. The values of the ratio in (19) for the 4th order example given by (17) are shown in Table 2.

CONCLUSIONS

The Routh Approximation method is a valuable technique for reducing the order of a linear dynamic system. An important advantage of the Routh approximation is that it is straightforward to use and is particularly well-suited for implementation on a digital computer. It should be noted that the Routh approximation is not based on computing the eigenvalues (poles of the transfer function) of the high-order model or performing a partial fraction expansion of the transfer function. This is a major advantage since determination of the eigenvalues (poles of the transfer function) of a high-order system is always difficult and often practically impossible.

The discussion of the Routh approximation in this paper has been restricted to single input-output, linear, fixed (time-invariant) dynamic systems for which the transfer function H(s) is given as a rational function of the complex frequency s. It has been further assumed that the high-order system described by H(s) is observable and controllable (no pole-zero cancellations occur) and also that the system is asymptotically stable (all poles have negative real parts). Although this simple class of linear systems forms a useful starting point to introduce the Routh approximation and study the fundamental properties of the method, it would appear to be too restrictive a class of linear systems for the majority of practical problems. Furthermore, the features of the Routh approximation indicate that the method is capable of being extended to handle a much broader class of linear systems, including

- Multiple input-output systems.
- Unstable systems.
- Distributed parameter systems.
- Discrete-time systems.
- Systems modeled by differential equations (the transfer function not being given).

Extensions of the Routh approximation method to handle the above systems is currently under investigation. A computer program is also being developed to facilitate application of the Routh approximation method and to study properties of the method.

REFERENCES

1. Chen, C. F., and Shieh, L.S., "A Novel Approach to Linear Model Simplification", International J. of Control, Vol. 8, No. 6, December 1968, pp. 561-570.

2. Schwarz, R.J., and Friedland, B., Linear Systems, McGraw-Hill Book Co., New York, 1965.

3. Astrom, K. J., Introduction to Stochastic Control Theory, Academic Press, New York and London, 1970.

Table 1

Approximate Transfer Functions Obtained by the Routh Approximation for a Fourth Order Example

$$H_1(s) = \frac{7.87883}{s + 1.19563}$$

$$H_2(s) = \frac{10.2695}{s + 1.78678} + \frac{3.00664}{s + 3.64140}$$

$$H_3(s) = \frac{10.0008}{s + 1.74893} + \frac{4.01567}{s + 4.82318} + \frac{0.405114}{s + 18.1243}$$

$$H(s) = H_4(s) = \frac{10.}{s + 1.74897} + \frac{4.}{s + 4.80153} + \frac{0.5}{s + 25.7783} + \frac{0.2}{s + 64.6712}$$

Table 2

Ratio of the Impulse Response Energy for the k-th Order Approximation to the Total Impulse Response Energy for the 4th Order Example

Order of the Approximation	Ratio
1st	60.15%
2nd	97.86%
3rd	99.97%
4th	100.%

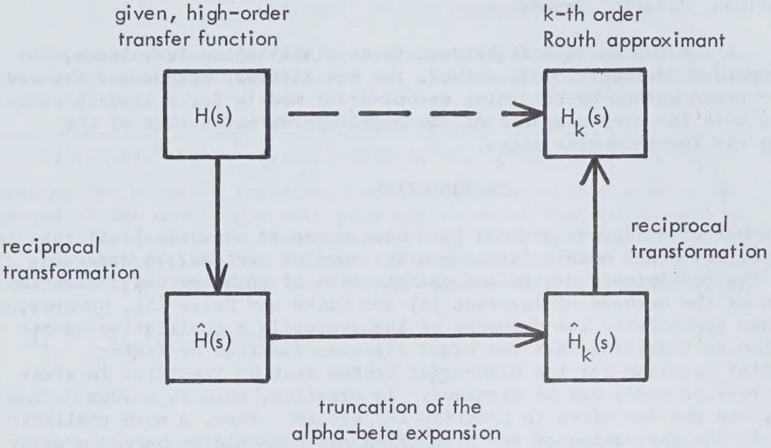

given, high-order transfer function
k-th order Routh approximant
H(s)
$H_k(s)$
reciprocal transformation
reciprocal transformation
$\hat{H}(s)$
$\hat{H}_k(s)$
truncation of the alpha-beta expansion

Figure 1 Routh Approximation Method

OPTIMUM APPROXIMATION OF HIGH-ORDER SYSTEMS BY LOW-ORDER MODELS USING RECENT GRADIENT METHODS

J.W. BANDLER, N.D. MARKETTOS and N.K. SINHA
Department of Electrical Engineering and
Communications Research Laboratory
McMaster University
Hamilton, Ontario, Canada

Abstract A comparison is made between three minimization techniques, the highly-regarded Fletcher-Powell method, the new Fletcher method and the new Jacobson-Oksman method by obtaining second-order models for a seventh-order system so that the step-response of the model approximates that of the system in the least-squares sense.

INTRODUCTION

A number of different methods have been proposed recently [1-5] for determining low-order models for high-order complex systems, in order to simplify the preliminary design and optimization of such systems. With the exception of the methods of Anderson [2] and Sinha and Pille [5], however, all of them approximate the response of the system in a qualitative manner in addition to requiring that the exact transfer function or vector differential equation for the high-order system must be specified in order that the reduced model may be obtained. In practice, this is seldom possible, and one has often to identify the system. Thus, a more realistic approach to the derivation of the simplified model should be based on using directly the measured input-output data for the system. Furthermore, it may be desirable to obtain an optimum model of a given order minimizing a specified criterion of error between the response of the model and that of the high-order system to a given input function, usually a unit step. Some suitable criteria for obtaining such a model are the minimization of (i) the sum of the squares of the error between sampled values of the two responses [2,5], (ii) the sum of the absolute values of the sampled errors, (iii) the sum of the pth power of the sampled errors, (iv) the maximum value of the errors between the responses, and (v) the maximum value of the perpendicular distance between the responses [6].

In a recent paper [7], Sinha and Bereznai have used the pattern search method of Hooke and Jeeves [8] for the determination of optimum low-order models using various criteria, and have calculated a number of such models for a given seventh-order system.

Due to the poor convergence properties of the pattern search algorithm, it was decided to investigate the application of efficient gradient methods to this problem. As well as the highly-regarded Fletcher-Powell method [9], two new gradient methods, one by Fletcher [10] and the other by Jacobson and Oksman [11] are available. This paper, therefore, attempts a comparative study of these three gradient methods by using each of them to obtain optimum second-order models, based on the least-squares criterion, for the same seventh-order system that was approximated by Sinha and Bereznai [7].

This work was supported by grants A7239 and A3374 from the National Research Council of Canada.

The statement of the problem will be followed by a brief review of the two new gradient methods. The results of computation and a critical comparison between the three methods will then be presented.

It may be added that as far as the authors are aware, this is the first comparative study of these three gradient methods.

STATEMENT OF THE PROBLEM

Given the following transfer function of a seventh-order system, representing the control-system of the pitch rate of a supersonic transport aircraft [12]

$$G(s) = \frac{375000(s+0.08333)}{s^7+83.64s^6+4097s^5+70342s^4+853703s^3+2814271s^2+3310875s+281250} \tag{1}$$

determine the transfer function, H(s), of a second-order model such that the response of the model to a unit step approximates that of the system in the least-squares sense. In other words, if $g_1(t)$ is the response of G(s) to a unit step and $h_1(t)$ is the response of the model H(s) to a unit step then the parameters of the model must be selected such that the objective function

$$J_1 = \int_0^\infty [g_1(t) - h_1(t)]^2 dt \tag{2}$$

is minimized.

As it is desired to solve the problem using a digital computer, the objective function (2) can be conveniently replaced by the following discrete version

$$J_2 = \sum_{k=1}^{N} [g_1(kT) - h_1(kT)]^2 \tag{3}$$

where N is sufficiently large to include the portion of response which should be approximated, and T a suitable sampling interval. As the system takes a considerable time to reach the steady-state value, and since it is more desirable to emphasize the initial portion of the transient response, it was decided to select the parameters of the second-order model so that its steady-state response was constrained to be equal to the steady-state response of the system. The application of this constraint made it possible to reduce the value of N without increasing the error for large values of t.

The resulting second-order models were of two types, those with a finite zero, and those without a finite zero. These may be written, respectively, as

$$H_1(s) = \frac{Ea_0}{s^2+a_1s+a_0} \tag{4}$$

and

$$H_2(s) = \frac{b_1s+Ea_0}{s^2+a_1s+a_0} \tag{5}$$

where E is the steady-state response of the system to a unit step input. The model represented by equation (4) has two variable parameters, while that represented by equation (5) has three variable parameters.

THE NEW FLETCHER METHOD

To give a basic description of the new Fletcher method it is necessary to briefly review the Fletcher-Powell method first. Suppose that it is desired to minimize a function $F(\underset{\sim}{x})$ of an n-dimensional vector $\underset{\sim}{x}$, given by

$$\underset{\sim}{x} = [x_1 \quad x_2 \; . \; . \; . \; x_n]^T \tag{6}$$

where the superscript T stands for transpose.

Let $\underset{\sim}{g}(\underset{\sim}{x})$ be the gradient of $F(\underset{\sim}{x})$ with respect to $\underset{\sim}{x}$ and let $\underset{\sim}{G}$ denote the corresponding Hessian matrix. The inverse Hessian $\underset{\sim}{G}^{-1}$ will be approximated by the matrix $\underset{\sim}{H}$. Define

$$\underset{\sim}{\delta} = \Delta\underset{\sim}{x} \tag{7}$$

as the increment in $\underset{\sim}{x}$, which is the correction made on the parameter vector $\underset{\sim}{x}$ in order to decrease $F(\underset{\sim}{x})$.

The main feature of the Fletcher-Powell method is that the increment

$$\underset{\sim}{\delta} = \alpha\underset{\sim}{s} \tag{8}$$

is taken along the direction

$$\underset{\sim}{s} = -\underset{\sim}{H}\underset{\sim}{g} \tag{9}$$

where α in (8) is that value of λ which minimizes $F(\underset{\sim}{x} + \lambda\underset{\sim}{s})$ along the direction of $\underset{\sim}{s}$. In practice, α is determined by a linear search. The matrix $\underset{\sim}{H}$ is updated at each iteration using the formula

$$\underset{\sim}{H}^{i+1} = \underset{\sim}{H}^{i} + \frac{\underset{\sim}{\delta}^{i}\underset{\sim}{\delta}^{i^T}}{\underset{\sim}{\delta}^{i^T}\underset{\sim}{\chi}^{i}} - \frac{\underset{\sim}{H}^{i}\underset{\sim}{\chi}^{i}\underset{\sim}{\chi}^{i^T}\underset{\sim}{H}^{i}}{\underset{\sim}{\chi}^{i^T}\underset{\sim}{H}^{i}\underset{\sim}{\chi}^{i}} \tag{10}$$

where

$$\underset{\sim}{\chi}^{i} = \underset{\sim}{g}^{i+1} - \underset{\sim}{g}^{i} \tag{11}$$

and the superscript i denotes the value at the ith iteration.

The updating formula (10) has the property that if $\underset{\sim}{H}^i$ is positive definite then $\underset{\sim}{H}^{i+1}$ must also be positive definite. Since $\underset{\sim}{H}$ is initially the unit matrix, $\underset{\sim}{H}^i$ is positive definite for all i.

The method has the property of quadratic convergence, that is, for a quadratic function the minimum can be located in at most n iterations. However, it depends on accurate location of the minimum along each direction of search. This is done by cubic interpolation which, although it is the most efficient method of search, requires several function and gradient evaluations.

The new Fletcher method dispenses with the linear search. The property of quadratic convergence, which depends on linear search, is replaced by a property which requires, for quadratic functions, that the eigenvalues of $\underset{\sim}{H}$ tend monotonically to those of $\underset{\sim}{G}^{-1}$. The decrease in F must be sufficiently large to guarantee ultimate convergence. This is taken care of by the following test. The change, ΔF in F on an iteration would be expected by Taylor's series expansion to be approximately $\underset{\sim}{g}^T\underset{\sim}{\delta}$ for a small $\underset{\sim}{\delta}$, but much less than $\underset{\sim}{g}^T\underset{\sim}{\delta}$ in absolute value when the position of the minimum along a line is overestimated. The change in F relative to $\underset{\sim}{g}^T\underset{\sim}{\delta}$ cannot become arbitrarily small if

$$\frac{\Delta F}{g^T \delta} \geq \mu \qquad (12)$$

where $0<\mu<<1$, a preassigned small quantity set at 0.0001. If corrections are determined by

$$\delta = -\lambda H g \qquad (13)$$

then trying values of $\lambda=1,w,w^2,w^3,\ldots$ for $w=0.1$ will eventually produce a δ that satisfies equation (12).

Although the conditions imposed so far are simple, failure can occur because H, updated using formula (10), can become singular. To overcome this problem the following updating formula is derived

$$H^{i+1} = H^i - \frac{\delta^i \gamma^{i^T} H^i}{\delta^{i^T} \gamma^i} - \frac{H^i \gamma^i \delta^{i^T}}{\delta^{i^T} \gamma^i} + \left(1 + \frac{\gamma^{i^T} H^i \gamma^i}{\delta^{i^T} \gamma^i}\right) \frac{\delta^i \delta^{i^T}}{\delta^{i^T} \gamma^i} \qquad (14)$$

Denote H^{i+1} in (14) as H_1^{i+1} and H^{i+1} in (10) as H_2^{i+1}. The formula in (14) has the property that the eigenvalues of H tend monotonically to those of G^{-1}. Note, however, that H_1^{i+1} does not necessarily replace H_2^{i+1}, as one of the two is chosen for updating on the basis of the following test. If

$$\delta^{i^T} \gamma^i \geq \gamma^{i^T} H^i \gamma^i \qquad (15)$$

then H_1^{i+1} is used; otherwise H_2^{i+1} is used.

The algorithm is terminated when $||\delta|| < \varepsilon$, a tolerance level, which was set at 1.0×10^{-6}.

THE JACOBSON-OKSMAN METHOD

This method is based on homogeneous rather than quadratic functions. A consequence of this is that convergence is obtained in n+2 steps for a homogeneous function of the form

$$F(x) = \frac{1}{\theta}(x-\check{x})^T g(x) + F(\check{x}) \qquad (16)$$

where x is an n-dimensional parameter vector as before, $g(x)$ is the gradient of $F(x)$ as defined previously, θ is the degree of homogeneity and $\check{x}$ is the location of the minimum of $F(x)$. On the other hand, a quadratic objective function may be expressed as

$$F(x) = \tfrac{1}{2}(x-\check{x})^T Q(x-\check{x}) + F(\check{x}) \qquad (17)$$

where Q is a constant positive definite matrix. Thus we see that equation (17) is a special case of the homogeneous equation (16) with $\theta = 2$.

The basis of the method will now be discussed. Equation (16) may be rearranged as

$$\check{x}^T g(x) + \theta\, F(x) - w = x^T\, g(x) \qquad (18)$$

where

$$w = \theta\, F(\check{x}) \; . \qquad (19)$$

Let

$$v \triangleq \underset{\sim}{x}^T \underset{\sim}{g}(\underset{\sim}{x})$$

$$\underset{\sim}{y} \triangleq [\underset{\sim}{g}^T(\underset{\sim}{x}) \quad F(\underset{\sim}{x}) \quad -1]^T \tag{20}$$

$$\underset{\sim}{\alpha} \triangleq [\check{\underset{\sim}{x}}^T \quad \theta \quad w]^T$$

For some point $\underset{\sim}{x}_{i+1}$, where in the interest of clarity, subscripts now denote iteration numbers, equation (18) becomes

$$v_{i+1} = \underset{\sim}{\alpha}^T \underset{\sim}{y}_{i+1} = \underset{\sim}{y}_{i+1}^T \underset{\sim}{\alpha} \tag{21}$$

where $\underset{\sim}{\alpha}$ and $\underset{\sim}{y}$ are (n+2)-vectors. The vector $\underset{\sim}{\alpha}$ contains the unknowns $\check{\underset{\sim}{x}}$ and w and must be determined. If we evaluate v and $\underset{\sim}{y}$ at n+2 distinct points, $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \cdots, \underset{\sim}{x}_{n+2}$, so that the resultant $\underset{\sim}{y}_i$ are linearly independent, we have

$$\begin{bmatrix} \underset{\sim}{y}_1^T \\ \underset{\sim}{y}_2^T \\ \vdots \\ \underset{\sim}{y}_{n+2}^T \end{bmatrix} \underset{\sim}{\alpha} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n+2} \end{bmatrix} \tag{22}$$

or, in matrix form

$$\underset{\sim}{Y} \underset{\sim}{\alpha} = \underset{\sim}{v} \tag{23}$$

Since the $\underset{\sim}{y}_i$ are linearly independent, the matrix $\underset{\sim}{Y}$ is non-singular, giving

$$\underset{\sim}{\alpha} = \underset{\sim}{Y}^{-1} \underset{\sim}{v} \tag{24}$$

To avoid matrix inversion, a recursive formula is used as new $\underset{\sim}{y}_i$ and v_i are evaluated. This is done as follows:

Let $\underset{\sim}{P}_0^{-1} = \underset{\sim}{I}$ an (n+2)x(n+2) identity matrix

and

$\underset{\sim}{v}_0 = \underset{\sim}{\alpha}_0$, where $\underset{\sim}{\alpha}_0$ is given.

At each iteration we successively replace corresponding rows and elements of $\underset{\sim}{P}_0^{-1}$ and $\underset{\sim}{v}_0$ with calculated values of $\underset{\sim}{y}_i$ and v_i, respectively, using

$$\underset{\sim}{P}_{i+1}^{-1} = \underset{\sim}{P}_i^{-1} + \underset{\sim}{e}_j (\underset{\sim}{y}_{i+1}^T - \underset{\sim}{e}_j^T \underset{\sim}{P}_i^{-1}) \tag{25}$$

$$\underset{\sim}{v}_{i+1} = \underset{\sim}{v}_i + \underset{\sim}{e}_j (v_{i+1} - \underset{\sim}{e}_j^T \underset{\sim}{v}_i) \tag{26}$$

where $\underset{\sim}{e}_j$ is a unit vector having unity as the jth element and zeros elsewhere, with j = i+1.

Using Householder's formula

$$\underset{\sim}{P}_{i+1} = \underset{\sim}{P}_i - \frac{\underset{\sim}{P}_i \underset{\sim}{e}_j (\underset{\sim}{y}_{i+1}^T \underset{\sim}{P}_i - \underset{\sim}{e}_j^T)}{\underset{\sim}{y}_{i+1}^T \underset{\sim}{P}_i \underset{\sim}{e}_j} \tag{27}$$

Successive estimates of the vector $\underset{\sim}{\alpha}$ are given by

$$\underset{\sim}{\alpha}_{i+1} = \underset{\sim}{\alpha}_i + \frac{\underset{\sim}{P}_i \; \underset{\sim}{e}_j (v_{i+1} - y_{i+1}^T \underset{\sim}{\alpha}_i)}{\underset{\sim}{y}_{i+1}^T \underset{\sim}{P}_i \underset{\sim}{e}_j} \tag{28}$$

Given the linear independence of $\underset{\sim}{y}_i$ it can be shown that for a homogeneous function

$$\underset{\sim}{P}_{n+2} = \underset{\sim}{Y}^{-1} \tag{29}$$

and

$$\underset{\sim}{\alpha}_{n+2} = \underset{\sim}{\alpha} \tag{30}$$

Thus, the algorithm finds the minimum, the degree of homogeneity and the value of the minimum after n+2 iterations.

APPLICATION TO THE PROBLEM OF OPTIMUM APPROXIMATION

The gradient methods discussed in the previous two sections depend considerably upon the availability of the first partial derivatives of the objective function with respect to the parameters of the model. Fortunately, it is fairly easy to derive analytical expressions for the gradient for a second-order model. Hence, the methods could be applied directly.

Since the efficiency of an optimization method is the speed with which the algorithm proceeds to the optimum, a fair indication of performance is the number of times the function and its gradient have to be evaluated. This is a valid assumption for the problem under consideration since the time taken for one evaluation of the function and its gradient is much more than the "housekeeping"-logic and other simple operations in the algorithm. Moreover, in the following discussion, a function evaluation will imply the complete evaluation of the function and its gradient, since only gradient methods are being considered.

RESULTS

First the two-parameter problem resulting from equation (4) was tried. The computer used was a CDC 6400 and typical C.P. times for the two-parameter problem were about 4 seconds. Three different starting points were considered for each of the three methods and in every case the algorithms ultimately converged to the same optimum parameters a_0=3.195912, a_1=2.281056, with the optimum value of the objective function 7.50758 x 10^{-4} and the components of the gradient less than 1.0 x 10^{-9}. Figure 1 shows the corresponding response. Table 1 compares the number of function evaluations required for each method for the objective function to reach the value 7.50759 x 10^{-4}. This value is 1.0 x 10^{-9} higher than the optimum ultimately obtained. From Table 1 it can be seen that both the new Fletcher method and the Jacobson-Oksman method show an improvement over the Fletcher-Powell method with the Jacobson-Oksman method being slightly better than the new Fletcher method.

For the three-parameter problem represented by equation (5) typical computer C.P. times were about 6 seconds. Again three different starting points were tried with each method and in each case the algorithms converged to the optimum a_0=1.997397, a_1=1.660663, b_1=4.370715 x 10^{-2} with the optimum value of the objective function 1.582215 x 10^{-4} and the gradient components less than 1.0 x 10^{-9}. Figure 2 shows the corresponding response. Table 2 compares the number of function evaluations required to reach the objective function value 1.582225 x 10^{-4} which is 1.0 x 10^{-9} higher than the optimum.

Starting points	$a_0=3.0$, $a_1=2.0$	$a_0=0.5$, $a_1=0.5$	$a_0=1.0$, $a_1=1.0$
Jacobson-Oksman	21	19	14
New Fletcher	21	22	19
Fletcher-Powell	29	49	32

Table 1: Number of function evaluations required to reach the objective function value 7.50759 x 10^{-4} for the 2-parameter problem.

Here again it is seen that the Jacobson-Oksman and the new Fletcher methods are superior to the Fletcher-Powell method and of the two methods the Jacobson-Oksman is slightly better. However, there was one case, not shown, when the Jacobson-Oksman method failed. This was due to a very large step that made the objective function too large for the computer to handle with single precision. In situations like this one is tempted to limit the step size, however it was decided not to interfere with the algorithm in any way, as this might slow down the method at the beginning.

Starting points	$a_0=1.0$, $a_1=1.0$, $b_1=1.0$	$a_0=0.5$, $a_1=2.0$, $b_1=4.0$	$a_0=1.0$, $a_1=0.5$, $b_1=0.1$
Jacobson-Oksman	39	39	29
New Fletcher	27	76	35
Fletcher-Powell	60	274	58

Table 2: Number of function evaluations required to reach the objective function value 1.582225 x 10^{-4} for the 3-parameter problem.

CONCLUSIONS

From the example considered, it may be concluded that all the three gradient methods can be successfully used for the determination of optimum low-order models for a high-order system. However, the two new methods are much superior to the older but highly-regarded Fletcher-Powell method. Between the two new methods, there is not much to choose, although the Jacobson-Oksman method was slightly faster for most of the cases considered.

ACKNOWLEDGEMENT

The authors are particularly grateful to Dr. D.H. Jacobson and Dr. R. Fletcher for making available their computer programs.

REFERENCES

[1] E.J. Davison, "A method for simplifying linear dynamic systems", IEEE Trans. Automatic Control, vol. AC-11, pp. 93-101, January 1966.

[2] J.H. Anderson, "Geometrical approach to reduction of dynamic systems", Proc. IEE, vol. 114, pp. 1014-1018, July 1967.

[3] C.F. Chen and L.S. Shieh, "A novel approach to linear model simplification", Preprints of the 1968 Joint Automatic Control Conference, Ann Arbor, Mich., pp. 454-460, June 1968.

[4] D. Mitra, "The reduction of complexity of linear time-invariant dynamic systems", Proc. 4th IFAC Congress, Warsaw, Poland, pp. 19-33, June 1969.

[5] N.K. Sinha and W. Pille, "A new method for reduction of dynamic systems", International Journal of Control, vol. 14, pp. 111-118, July 1971.

[6] G.T. Bereznai and N.K. Sinha, "A new minimax objective for automated system design", Electronics Letters, vol. 6, pp. 847-849, December 1970.

[7] N.K. Sinha and G.T. Bereznai, "Optimum approximation of high-order systems by low-order models", International Journal of Control, to be published in December 1971.

[8] R. Hooke and T.A. Jeeves,"'Direct search' solution of numerical and statistical problems, J.ACM, vol. 8, pp. 212-219, April 1961.

[9] R. Fletcher and M.J.D. Powell, "A rapidly convergent descent method for minimization", Computer J., vol. 6, pp. 163-168, June 1963.

[10] R. Fletcher, "A new approach to variable metric algorithms", Computer J., vol. 13, pp. 317-322, August 1970.

[11] D.H. Jacobson and W. Oksman, "An algorithm that minimizes homogeneous functions of n variables in n+2 iterations and rapidly minimizes general functions", Division of Engineering and Applied Physics, Harvard University, Technical Report No. 618, October 1970.

[12] R.C. Dorf, Modern Control Systems. Reading, Mass.: Addison-Wesley, 1967, pp. 167-168.

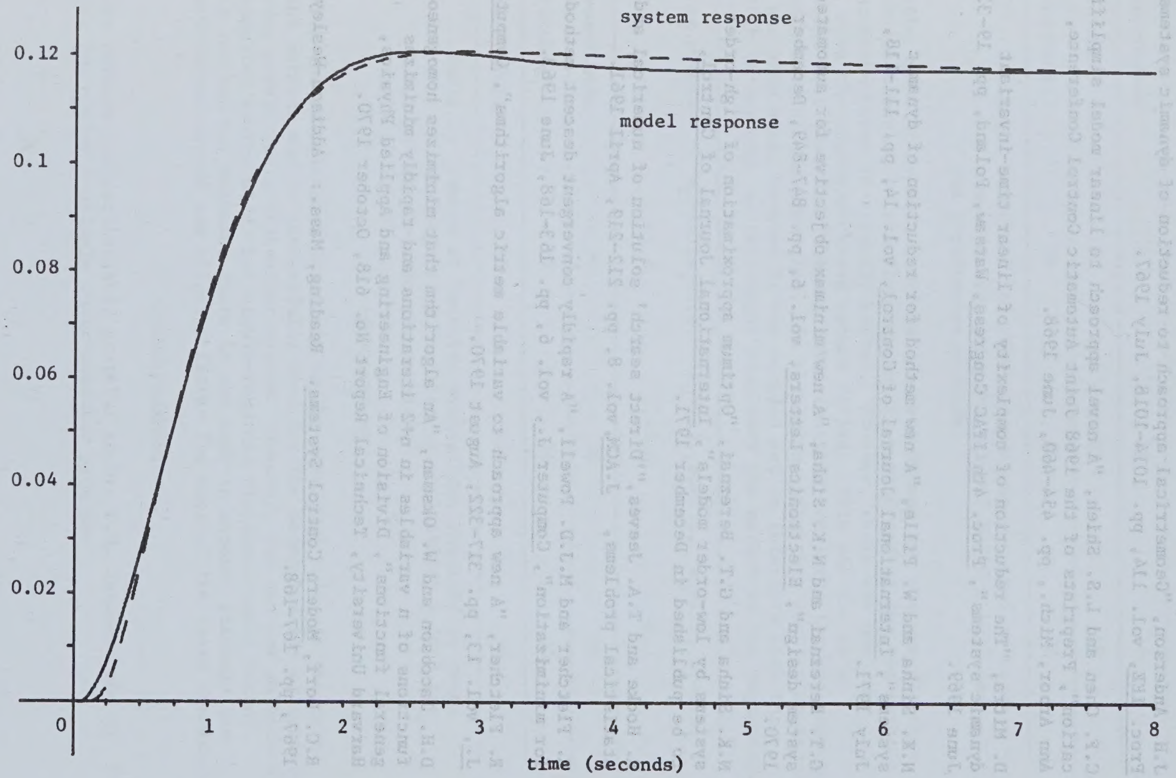


Figure 1. Response of the 2-parameter model.

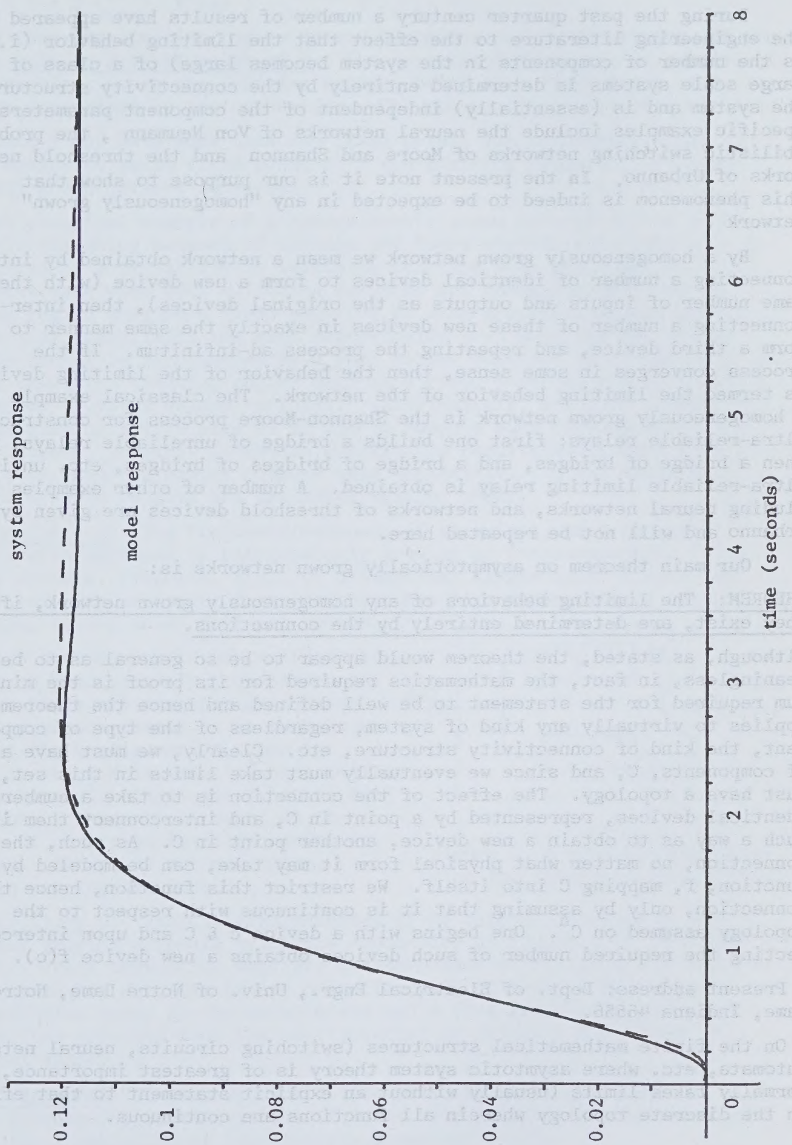


Figure 2. Response of the 3-parameter model.

ON ASYMPTOTIC SYSTEM THEORY: AN APPROACH TO SOME LARGE SCALE SYSTEMS

CELSO DE RENNA E SOUZA & RICHARD SAEKS
COPPE/UFRJ
Rio de Janeiro, GB, Brazil*

During the past quarter century a number of results have appeared in the engineering literature to the effect that the limiting behavior (i.e. as the number of components in the system becomes large) of a class of large scale systems is determined entirely by the connectivity structure of the system and is (essentially) independent of the component parameters. Specific examples include the neural networks of Von Neumann , the probabilistic switching networks of Moore and Shannon and the threshold networks of Urbanno. In the present note it is our purpose to show that this phenomenom is indeed to be expected in any "homogeneously grown" network .

By a homogeneously grown network we mean a network obtained by interconnecting a number of identical devices to form a new device (with the same number of inputs and outputs as the original devices), then interconnecting a number of these new devices in exactly the same manner to form a third device, and repeating the process ad-infinitum. If the process converges in some sense, then the behavior of the limiting device is termed the limiting behavior of the network. The classical example of a homogeneously grown network is the Shannon-Moore process for constructing ultra-reliable relays: first one builds a bridge of unreliable relays, and then a bridge of bridges, and a bridge of bridges of bridges, etc. until an ultra-reliable limiting relay is obtained. A number of other examples including neural networks, and networks of threshold devices are given by Urbanno and will not be repeated here.

Our main theorem on asymptotically grown networks is:

THEOREM: The limiting behaviors of any homogeneously grown network, if they exist, are determined entirely by the connections.

Although, as stated, the theorem would appear to be so general as to be meaningless, in fact, the mathematics required for its proof is the minimum required for the statement to be well defined and hence the theorem applies to virtually any kind of system, regardless of the type of component, the kind of connectivity structure, etc. Clearly, we must have a set of components, C, and since we eventually must take limits in this set, it must have a topology. The effect of the connection is to take a number of identical devices, represented by a point in C, and interconnect them in such a way as to obtain a new device, another point in C. As such, the connection, no matter what physical form it may take, can be modeled by a function, f, mapping C into itself. We restrict this function, hence the connection, only by assuming that it is continuous with respect to the topology assumed on C[Δ]. One begins with a device c ε C and upon interconnecting the required number of such devices obtains a new device f(c).

---

* Present address: Dept. of Electrical Engr., Univ. of Notre Dame, Notre Dame, Indiana 46556.

[Δ] On the finite mathematical structures (switching circuits, neural nets, automata, etc. where asymtotic system theory is of greatest importance, one normally takes limits (usually without an explicit statement to that effect) in the discrete topology wherein all functions are continuous.

Repeating the process yields $f(f(c)) \triangleq f^2(c)$ and in general after n iterations we have $f^n(c)$. The limiting device is therefore given by $c_o = \lim_{n\to\infty} f^n(c)$ where the limit is taken in the topology of C. Finally, since f is continuous in this topology we have

$$f(c_o) = f\left(\lim_{n\to\infty} f^n(c)\right) = \lim_{n\to\infty} f^{n+1}(c) = c_o$$

showing that any limiting regime of the homogeneously grown network, if any exists, must be a fixed point of f. The function f, however, is a model of the connection; hence any limiting regime, being some fixed point of f, is entirely determined by the connection. The theorem is, therefore, verified.

Since f may have more than one fixed point, the component value, c, may determine to which fixed point of f the limiting regime of the homogeneously grown network converges.

The classical example of a homogeneously grown network is the probabilistic switching network of Shannon and Moore wherein the connection is represented by a probabilistic switching function h(p) and, indeed, it is well known that when such switching networks are homogenously grown the limiting regime must be a fixed point of h. The present theorem, however, assures that essentially the same result holds even when the connection function cannot be represented by a real valued function of a real variable. In that case, the limiting regime must be a fixed point of a system valued function of a system valued variable. Unfortunately, standard fixed point theorems give us little clue as to how to compute such fixed points. It is often, however, possible to deduce qualitative results about such fixed points even when one cannot compute them explicitly. In the sequel two examples of such results are described without proof.

Corollary 1: The frequency response, if it exists, of a limiting linear system obtained by homogeneously growing an interconnection (with adders, scalars, feedback, etc.) of n identical single-input single-output devices with arbitrary frequency response is a superposition of step functions, taking on at most n+1 values.

As a second example we consider the problem of homogeneously grown automata. Although there are many possible variations on such a theme, as an example we consider the configuration shown in figure 1, where M is taken to be a Moore-type machine with input set X, state set S and output taken to be equal to the state. Of course we identify equivalent machines. Here Δ is any memoryless function mapping SxS into S.

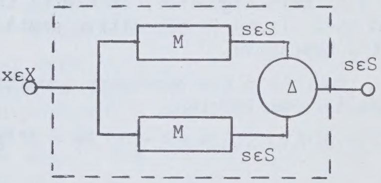


Figure 1. A possible configuration for interconnected automata

Corollary 2: The limiting machine, if it exists, obtained by homogeneously growing automata using the connection of figure 1 is "Δ-linear" (i.e. its state transition maps are endomorphisms of the groupoid defined on the state set S by the binary operation Δ.) Of course, if our state set is an additive group in the usual sense and Δ is a group adder, the limiting machines obtained from the configuration of figure 1 are linear in the usual sense.

Acknowledgement: The authors would like to thank their colleague at COPPE, Professor Franco Preparata, for the many helpful discussions which led to the formulation of the concepts presented herein.

ASYMPTOTIC EXPANSIONS OF SINGULARLY PERTURBED QUASI-LINEAR OPTIMAL SYSTEMS

PEDDAPULLAIAH SANNUTI
Rutgers University, New Brunswick, New Jersey

ABSTRACT

A class of two point boundary value problems (TPBVP) which arise in fixed final time free end point optimal control problems are considered. An asymptotic power series solution of the TPBVP is constructed with respect to a parameter whose perturbation changes the order of the problem. Based on a stability hypothesis, the proof of asymptotic correctness is accomplished through a successive approximation scheme.

1. PROBLEM STATEMENT

Let us consider a quasi-linear system,

$$\dot{x} = g_1(x,t) + B_1(t)z + C_1(t)u \quad , \quad x(0) = x_0 \tag{1.1a}$$

$$\lambda\dot{z} = G_1(x,t) + B_2(t)z + C_2(t)u \quad , \quad z(0) = z_0 \tag{1.1b}$$

where x and z are n- and m-dimensional state vectors respectively, u is an r-dimensional control vector and λ is a nonnegative scalar parameter. Note for $\lambda = 0$, z ceases to be a state vector and the order of system (1.1) reduces to n. The objective of the design is to find an optimal control $u(t,\lambda)$ which takes the initial states x_0 and z_0 to the free end, with fixed final time, while minimizing the performance index,

$$J = \int_0^T (V(x,t) + \frac{1}{2} u' R(t)u)dt. \tag{1.2}$$

The functions g_1, G_1, B_1, B_2, C_1, C_2, V and R could depend on the parameter λ. It is assumed that these functions are infinitely differentiable in all their arguments, that R(t) is a symmetric positive definite matrix and that V(x,t) is a positive semidefinite scalar function. The prime denotes transpose.

To obtain the necessary conditions for an optimum control, we introduce the Hamiltonian,

$$H(x,z,p,q,u,\lambda,t) = -V - \frac{1}{2}u' Ru + p'(g_1 + B_1 z + C_1 u) + q'(G_1 + B_2 z + C_2 u) \tag{1.3}$$

where p and q are the costate variables corresponding to x and z respectively. The maximum principle implies that along an optimal trajectory

$$\nabla_u H = - Ru + C_1'p + C_2'q = 0 \tag{1.4}$$

while the costate variables p and q satisy the equations

$$\dot{p} = - \nabla_x H = \nabla_x V(x,t) - g_{1x}'(x,t)p - G_{1x}'(x,t)q$$

$$\lambda\dot{q} = - \nabla_z H = - B_1'(t)p - B_2'(t)q \tag{1.5}$$

along with the final conditions

$$p(T) = 0 \text{ and } q(T) = 0. \tag{1.6}$$

From (1.4),

$$u(t) = R^{-1}(C_1'p + C_2'q). \tag{1.7}$$

Now we rewrite the state and costate equations (1.1) and (1.5),

$$\begin{aligned}
\dot{x} &= g(x,p,z,q,\lambda,t) = g_1 + S_1 p + B_1 z + Sq \\
\dot{p} &= f(x,p,q,\lambda,t) = \nabla_x V - g'_{1x} p - G'_{1x} q \qquad (1.8) \\
\lambda\dot{z} &= G(x,p,z,q,\lambda,t) = G_1 + S'p + B_2 z + S_2 q \\
\lambda\dot{q} &= F(p,q,\lambda,t) = -B_1'p - B_2'q
\end{aligned}$$

where S_1, S and S_2 are the matrices
$S_1(t) = C_1(t)R^{-1}(t)\ C_1'(t)$; $S(t) = C_1(t)R^{-1}(t)C_2'(t)$ and $S_2(t)=C_2(t)R^{-1}(t)C_2'(t)$.
The boundary conditions for (1.8) are

$$x(0) = x_0,\ p(T) = 0,\ z(0) = z_0 \text{ and } q(T) = 0. \qquad (1.9)$$

Equations (1.8) and (1.9) define a two point boundary value problem (TPBVP) of order $2(n+m)$. The order of the problem drops from $2(n+m)$ for $\lambda > 0$ to $2n$ for $\lambda = 0$, thus we have a singularly perturbed TPBVP. The reduced problem is given by

$$\dot{X}_0 = g(X_0,P_0,Z_0,Q_0,0,t) \qquad (1.10a)$$

$$\dot{P}_0 = f(X_0,P_0,Q_0,0,t) \qquad (1.10b)$$

$$0 = G(X_0,P_0,Z_0,Q_0,0,t) \qquad (1.10c)$$

$$0 = F(P_0,Q_0,0,t). \qquad (1.10d)$$

We can solve Z_0 and Q_0 from (1.10c) and(1.10d) and substitute them into (1.10a) and (1.10b), thus getting a 2n-th order problem with boundary conditions,

$$X_0(0) = x_0 \text{ and } P_0(T) = 0. \qquad (1.11)$$

Alternatively one could neglect λ in the state equations (1.1) directly and then construct the necessary conditions for the optimal control. The resulting boundary value problem will be the same as that obtained above.

It is well known that the computational difficulties in obtaining an iterative solution of a TPBVP increase exponentially as the order of the problem increases. Another hidden difficulty for (1.8) is that the standard methods of stepwise integration require the use of step sizes which are far below the value of λ. Because of this, for small λ the direct solution of (1.8) and (1.9) is computationally prohibitive. Based on intuition, one could probably neglect λ and obtain the control from the solution of the reduced problem (1.10) and (1.11). However, such a control alone may result in an unsatisfactory performance when applied to the full problem (1.1), since in general, $Z_0(0)$ of the reduced problem is completely different from the initial condition z_0 (see, for examples, [10]). Our interest in this paper is to determine the asymptotic expansions of the optimal control by determining the asymptotic solution of the full problem (1.8) in terms of the positive parameter λ as λ tends to zero. We will calculate all the terms in the expansions using only low-order systems. For a class of problems Sannuti and Kokotović gave earlier, a design procedure to calculate the first two terms of the expansions [6,11]. This procedure was based on the work of Vasil'eva [13] and Tupčiev [12]. Hadlock [3] has given another modification of Vasil'eva's and Tupčiev's results so that a general TPBVP can be treated. However, he has not explicitly given any method of choosing auxiliary conditions for some of the terms in the asymptotic solution. Further, he did not attempt to prove the validity of expansions. Recently, O'malley [9] has considered asymptotic expansions for a linear state regulator problem. He proves the asymptotic validity by utilizing a constructed set of linearly independent solutions of TPBVP. For nonlinear problems one can not rely on such a proof. The proof in this paper is basically an iterative scheme similar to those in Levin [8], Hadlock [4] and Hoppensteadt [5]. We follow closely Hadlock's method, who used it to study the existence and behavior of solutions of problems similar to (1.8).

2. CONSTRUCTION OF THE FORMAL EXPANSIONS

We shall be able to uniquely obtain the complete asymptotic solution of the problem (1.8) and (1.9) under the following assumptions.

(H1): The reduced two point boundary value problem (1.10) and (1.11) has a unique solution.

(H2): The matrix $B_2(t)$ is a stable matrix, i.e., all its m eigenvalues have negative real parts for each fixed t, $0 \leq t \leq T$.

Hypothesis H2 is crucial and plays several roles. It determines the solvability of algebraic equations of the type (1.10) for Z_0 and Q_0. It rules out the turning point behavior. As will be seen later, it defines the behavior of boundary layer solutions.

Based on previous experience, we seek a formal solution of (1.8) and (1.9) of the form

$$\begin{aligned} x(t) &= X(t,\lambda) + \lambda m_1(\tau,\lambda) + \lambda n_1(\sigma,\lambda) \\ p(t) &= P(t,\lambda) + \lambda m_2(\tau,\lambda) + \lambda n_2(\sigma,\lambda) \\ z(t) &= Z(t,\lambda) + m_3(\tau,\lambda) + n_3(\sigma,\lambda) \\ q(t) &= Q(t,\lambda) + m_4(\tau,\lambda) + n_4(\sigma,\lambda) \end{aligned} \tag{2.1}$$

where

$$\tau = \frac{t}{\lambda} \text{ and } \sigma = \frac{T-t}{\lambda} .$$

Here the "outer expansion" (X,P,Z,Q) will formally satisfy the system (1.8) throughout $0 < t < T$ and will have an asymptotic power series expansion there as $\lambda \to 0$,

$$X(t,\lambda) = \sum_{j=0}^{\infty} X_j(t)\lambda^j , \quad P(t,\lambda) = \sum_{j=0}^{\infty} P_j(t)\lambda^j, \quad Z(t,\lambda) = \sum_{j=0}^{\infty} Z_j(t)\lambda^j, Q(t,\lambda) = \sum_{j=0}^{\infty} Q_j(t)\lambda^j. \tag{2.2}$$

It should be mentioned that the outer expansion (2.2) alone will not, in general, satisfy the boundary conditions (1.9), however. The expansion $(\lambda m_1, \lambda m_2, m_3, m_4)$ represents the "boundary layer correction" at the initial point $t = 0$, and the expansion $(\lambda n_1, \lambda n_2, n_3, n_4)$ represents the "boundary layer correction" at the final point $t = T$. They will have asymptotic power series expansions as $\lambda \to 0$,

$$m_i(\tau,\lambda) = \sum_{j=0}^{\infty} m_{ij}(\tau)\lambda^j ; \quad n_i(\sigma,\lambda) = \sum_{j=0}^{\infty} n_{ij}(\sigma)\lambda^j ; \; i = 1 \text{ to } 4. \tag{2.3}$$

The boundary layer coefficients $m_{ij}(\tau)$ and $n_{ij}(\sigma)$ will be so chosen that (2.1) will satisfy the boundary conditions (1.9), that $m_{ij}(\tau)$ will tend to zero as the left boundary layer coordinate τ tends to infinity and that $n_{ij}(\sigma)$ will tend to zero as the right boundary layer coordinate σ tends to infinity,

$$m_{ij}(\tau) \to 0 \text{ as } \tau \to \infty, \; n_{ij}(\sigma) \to 0 \text{ as } \sigma \to \infty, \; i = 1 \text{ to } 4. \qquad (2.4)$$

Now taking x_0 and z_0 independent of λ, and neglecting $m_i(\tau)$ at $t = T$ and $n_i(\sigma)$ at $t = 0$, the boundary conditions (1.9) are equivalent to

$$X_0(0) = x_0 \quad , \quad P_0(T) = 0 \qquad (2.5a)$$

$$m_{3\,0}(0) = z_0 - Z_0(0) \quad , \quad n_{4\,0}(0) = 0 \qquad (2.5b)$$

$$X_j(0) = -m_{1\;j-1}(0) \quad , \quad P_j(T) = -n_{2\;j-1}(0) \text{ for all } j \geq 1 \qquad (2.5c)$$

$$m_{3j}(0) = -Z_j(0) \quad , \quad n_{4j}(0) = -Q_j(T) \text{ for all } j \geq 1. \qquad (2.5d)$$

Note (2.4) implies that during the open interval $0 < t < T$ the asymptotic solution will be determined by the outer expansion alone, i.e., for each integer $N \geq 0$

$$x(t) = \sum_{j=0}^{N} X_j(t)\lambda^j + o(\lambda^N), \; p(t) = \sum_{j=0}^{N} P_j(t)\lambda^j + o(\lambda^N)$$

$$z(t) = \sum_{j=0}^{N} Z_j(t)\lambda^j + o(\lambda^N), \; q(t) = \sum_{j=0}^{N} Q_j(t)\lambda^j + o(\lambda^N)$$

uniformly on each interval $\delta \leq t \leq T - \delta$ for any $\delta > 0$. As (2.5) shows, however, the initial values $m_{1j}(0)$ and $n_{2j}(0)$ must be calculated in order to determine the outer expansion.

The outer expansion is obtained by substituting (2.2) into (1.8) and equating formally the coefficients of like powers of λ^j. We have equality at $\lambda = 0$ since X_0, P_0, Z_0 and Q_0 is the solution of the reduced problem. Equating first order coefficients, X_1, P_1, Z_1 and Q_1 must satisfy the linear equations,*

$$\dot{X}_1 = g_{1x}(t)X_1 + S_1(t)P_1 + B_1(t)Z_1 + S(t)Q_1 + g_\lambda(t) \qquad (2.6a)$$

$$\dot{P}_1 = f_x(t)X_1 - g'_{1x}(t)P_1 - G'_{1x}(t)Q_1 + f_\lambda(t) \qquad (2.6b)$$

$$0 = G_{1x}(t)X_1 + S'(t)P_1 + B_2(t)Z_1 + S_2(t)Q_1 + G_\lambda(t) - \dot{Z}_0 \qquad (2.6c)$$

$$0 = -B_1'(t)P_1 - B_2'(t)Q_1 + F_\lambda(t) - \dot{Q}_0. \qquad (2.6d)$$

---

* $f_x(t)$ denotes $\frac{\partial f}{\partial x}$ calculated along the solution of the reduced problem, $f_x(t) = \frac{\partial f}{\partial x}(X_0,P_0,Q_0,0,t)$. Other derivatives are similarly defined.

Since B_2 is non singular, (2.6c) and(2.6d)can be solved for Z_1 and Q_1 so that (2.6a) and (2.6b) can be rewritten in the form

$$\dot{X}_1 = A_1(t)\, X_1 + A_2(t)\, P_1 + L_{10}(t) \tag{2.7a}$$

$$\dot{P}_1 = A_3(t)\, X_1 - A'_1(t)\, P_1 + L_{20}(t) \tag{2.7b}$$

where

$$A_1(t) = g_{1x}(t) - B_1(t)B_2^{-1}(t)G_{1x}(t)$$

$$A_2(t) = (B_1(t)B_2^{-1}(t)C_2(t) - C_1(t))\, R^{-1}(t)\, (B_1(t)B_2^{-1}(t)C_2(t) - C_1(t))'$$

$$A_3(t) = -\frac{d}{dx}\nabla_x H = f_x(t),$$

and the expressions for L_{10} and L_{20} can be similarly obtained from (2.6). It can be noted that A_2 is positive semidefinite. Since X_0, P_0, Z_0 and Q_0 is an optimal solution of the reduced problem, $A_3(t)$ is also positive semidefinite. The boundary conditions for (2.7) are given by (2.5c),

$$X_1(0) = -m_{10}(0)\ ,\ P_1(T) = -n_{20}(0)\ . \tag{2.8}$$

Equations (2.7) and (2.8) represent a well known two point boundary value problem [7]. With

$$P_1(t) = -\, K(t)X_1(t) + \Pi_1(t) \tag{2.9}$$

this problem is reducible to a final value problem for a n×n dimensional symmetric matrix Riccati equation

$$\dot{K} = -\, K\, A_1 - A_1'K + K\, A_2K - A_3\ ;\ K(T) = 0 \tag{2.10}$$

and an n-dimensional linear vector equation

$$\dot{\Pi}_1 = (K\, A_2 - A_1')\Pi_1 + K\, L_{10} + L_{20};\ \Pi_1(T) = -\, n_{20}(0). \tag{2.11}$$

Once the matrix K and the vector Π_1 are found the substitution of (2.9) in (2.7a) allows X_1 to be obtained uniquely as the solution of

$$\dot{X}_1 = (A_1 - A_2K)X_1 + A_2\Pi_1 + L_{10} \tag{2.12}$$

for the initial condition $X_1(0) = -\, m_{10}(0)$. Thus the variables X_1, P_1, Z_1 and Q_1 are uniquely obtained. In general, we find X_j, P_j, Z_j and Q_j by solving

$$\begin{aligned}
\dot{X}_j &= g_{1x}(t)X_j + S_1(t)P_j + B_1(t)Z_j + S(t)Q_j + R_{1j-1}(t)\\
\dot{P}_j &= f_x(t)X_j - g'_{1x}(t)P_j - G'_{1x}(t)Q_j + R_{2j-1}(t)\\
0 &= G_{1x}(t)X_j + S'(t)P_j + B_2(t)Z_j + S_2(t)Q_j + R_{3j-1}(t)\\
0 &= -\, B_1'(t)P_j - B_2'(t)Q_j + R_{4j-1}(t)
\end{aligned} \tag{2.13}$$

with boundary values given by (2.5c), where $R_{i\ j-1}(t)$, $i = 1$ to 4 are successively known polynomials in X_r, P_r, Z_r and Q_r for all $r \leq j-1$. Note (2.13) differs from (2.6) only in nonhomogeneous terms. Hence the solution is easy and requires solving only linear equations of the form (2.11) and (2.12). Thus the terms of the outer expansion can be generated recursively with only the boundary values $X_j(0)$ and $P_j(T)$ for $j \geq 1$ to be specified, (2.5c).

The following simple lemma will be repeatedly used to obtain the boundary layer terms.

Lemma 1:

Consider a linear time invariant system

$$\dot{\alpha} = - B \alpha + b \,\mathrm{Exf}(-\upsilon t) \tag{2.14}$$

where b is a m-vector, B is a m × m matrix with the real parts of all eigenvalues less than $-\mu$ for some $\mu > 0$, and let υ be a positive parameter. Then there exists a unique α_0 such that the solution $\alpha(t)$ of (2.14) with the auxiliary condition, $\alpha(0) = \alpha_0$, satisfies the inequality*

$$|\alpha(t)| \leq \underline{K}\, \mathrm{Exf}(-\upsilon t) \text{ for all } t \geq 0$$

where $\underline{K}$ is a bounded constant.

Proof:

$$\alpha(t) = \mathrm{Exf}(-Bt)\left[\alpha_0 + \int_0^\infty \mathrm{Exf}((B-I\upsilon)s)b\, ds + \int_\infty^t \mathrm{Exf}((B-I\upsilon)s)b\, ds\right]$$

Since B is a stable matrix the first integral in the parenthesis exists. Let it be $= M_0$, choose $\alpha_0 = - M_0$,

$$\alpha(t) = \int_\infty^t \mathrm{Exf}(B(s-t) - I\upsilon s)b\, ds.$$

Observing the inequality [1],

$|\mathrm{Exf}(B(s-t))| \leq K_0\, \mathrm{Exf}(-\mu(s-t))$ for all $s \geq t$, we have

$$|\alpha(t)| \leq \int_t^\infty K_1\, \mathrm{Exf}(-\mu(s-t) - \upsilon s)ds \leq \underline{K}\, \mathrm{Exf}(-\upsilon t).$$

Since the outer expansion (2.2) formally satisfies the system (1.8) and since the right boundary layer correction $n_i(\sigma)$ is asymptotically negligible near $t = 0$, (2.1) and (1.8) imply that the left boundary layer correction $m_i(\tau)$ satisfies the system of equations,

$$\begin{aligned}
\frac{dm_1}{d\tau} &= \hat{g}(m_1,m_2,m_3,m_4,\lambda,\tau) \\
\frac{dm_2}{d\tau} &= \hat{f}(m_1,m_2,m_4,\lambda,\tau) \\
\frac{dm_3}{d\tau} &= \hat{G}(m_1,m_2,m_3,m_4,\lambda,\tau) \\
\frac{dm_4}{d\tau} &= - \lambda B_1'(\lambda\tau)m_2 - B_2'(\lambda\tau)m_4
\end{aligned} \tag{2.15}$$

---

* The norm of a matrix or a vector A is denoted by $|A|$ and is the sum of the magnitudes of all the elements of A.

successively. Using lemma 1, $m_{4j}(0)$ can be selected such that $m_{4j}(\tau)$ decay exponentially. $m_{3j}(0)$ will assume the initial condition $- Z_j(0)$. Since $B_2(0)$ is a stable matrix it is evident that $m_{ij}(\tau)$, i = 1 to 4, can be uniquely defined and decay exponentially as τ tends to infinity. Now (2.5c) gives $X_{j+1}(0) = - m_{1j}(0)$.

The right boundary layer correction terms $n_i(\sigma)$ are determined in an analogous manner. Since $m_i(\tau)$ are negligible at the final point t = T, (2.1) and (1.8) imply that

$$\frac{dn_1}{d\sigma} = \tilde{g}\,(n_1,\, n_2,\, n_3,\, n_4,\, \lambda,\, \sigma)$$

$$\frac{dn_2}{d\sigma} = \tilde{f}\,(n_1,\, n_2,\, n_4,\, \lambda,\, \sigma) \tag{2.17}$$

$$\frac{dn_3}{d\sigma} = \tilde{G}\,(n_1,\, n_2,\, n_3,\, n_4,\, \lambda,\, \sigma)$$

$$\frac{dn_4}{d\sigma} = \lambda\, B_1'\,(T - \lambda\sigma)\, n_2 + B_2'\,(T - \lambda\sigma)\, n_4$$

where

$$\tilde{g} = g(X(t),\, P(t),\, Z(t),\, Q(t),\lambda,t) - g(X(t)+\lambda n_1(\sigma),P(t)+\lambda n_2(\sigma),Z(t) + n_3(\sigma),Q(t) + n_4(\sigma),\, \lambda,\sigma)$$

in which t should be replaced by $T - \lambda\sigma$, and $\tilde{f}$, $\tilde{G}$ are similarly defined.

Substituting asymptotic power series and equating like coefficients of λ^j we obtain a system of equations for $n_{ij}(\sigma)$. The lowest order terms will satisfy

$$\frac{dn_{10}}{d\sigma} = - B_1(T)n_{30} - S(T)n_{40}$$

$$\frac{dn_{20}}{d\sigma} = G_{1x}'(X_0(T),T)n_{40}$$

$$\frac{dn_{30}}{d\sigma} = - B_2(T)n_{30} - S_2(T)n_{40}$$

$$\frac{dn_{40}}{d\sigma} = B_2'(T)n_{40}.$$

The solution $n_{i0}(\sigma)$ should vanish as σ tends to infinity, further (2.5b) dictates $n_{40}(0) = 0$. Using lemma 1 we choose $n_{30}(0) = 0$. Thus

$$n_{i0}(\sigma) \equiv 0 \quad , \quad i = 1 \text{ to } 4.$$

Now (2.5c) gives $P_1(T) = 0$. This condition on $P_1(T)$ agrees with what has been derived earlier [11].

In general the higher order terms $n_{ij}(\sigma)$ satisfy

where

$$\hat{g} = g(X(\lambda\tau) + \lambda m_1(\tau),\ P(\lambda\tau) + \lambda m_2(\tau),\ Z(\lambda\tau) + m_3(\tau),\ Q(\lambda\tau) + m_4(\tau),\ \lambda,\ \lambda\tau)$$

$$- g(X(\lambda\tau),\ P(\lambda\tau),\ Z(\lambda\tau),\ Q(\lambda\tau),\ \lambda,\ \lambda\tau)$$

and $\hat{f}$, $\hat{G}$ are similarly defined.

Substituting the asymptotic power series for $m_i(\tau)$ into (2.15) and equating coeffieicents, we obtain a system of equations for the left boundary layer terms. In particular, the lowest order terms will satisfy the linear system,

$$\frac{dm_{10}}{d\tau} = B_1(0)m_{30} + S(0)m_{40}$$

$$\frac{dm_{20}}{d\tau} = -\ G'_{1x}(X_0(0),0)m_{40}$$

$$\frac{dm_{30}}{d\tau} = B_2(0)m_{30} + S_2(0)m_{40}$$

$$\frac{dm_{40}}{d\tau} = -\ B'_2(0)m_{40}\ .$$

Note from (2.5b), $m_{30}(0) = z_0 - Z_0(0)$. The other auxiliary conditions must be selected in such a way that $m_{i0}(\tau)$ tends to zero as τ tends to infinity. Using lemma 1, we choose $m_{40}(0) = 0$,

$$m_{40}(\tau) \equiv 0\ ,\ m_{30}(\tau) = \mathrm{Exf}(B_2(0)\tau)(z_0 - Z_0(0)),$$

$$m_{20}(\tau) \equiv 0\ ,\ \text{and}\ m_{10}(\tau) = B_1(0)B_2^{-1}(0)\ \mathrm{Exf}(B_2(0)\tau)(z_0 - Z_0(0)).$$

Thus $m_{i0}(\tau)$ are uniquely defined. Now (2.5c) gives

$$X_1(0) = -\ m_{10}(0) = -\ B_1(0)B_2^{-1}(0)(z_0 - Z_0(0)).$$

This initial condition on $X_1(0)$ agrees with what has been derived earlier [11] by another method.

In general $m_{ij}(\tau)$ are obtained from (2.15) by successively equating like coefficients of λ^j,

$$\begin{aligned}
\frac{dm_{1j}}{d\tau} &= B_1(0)m_{3j} + S(0)m_{4j} + M_{1\ j-1}(\tau)\\
\frac{dm_{2j}}{d\tau} &= -\ G'_{1x}(X_0(0),0)m_{4j} + M_{2\ j-1}(\tau)\\
\frac{dm_{3j}}{d\tau} &= B_2(0)m_{3j} + S_2(0)m_{4j} + M_{3\ j-1}(\tau)\\
\frac{dm_{4j}}{d\tau} &= -\ B'_2(0)m_{4j} + M_{4\ j-1}(\tau)
\end{aligned} \qquad (2.16)$$

where the exponentially decaying terms $M_{i\ j-1}(\tau)$, $i = 1$ to 4 are known

$$\begin{aligned}
\frac{dn_{1j}}{d\sigma} &= - B_1(T)n_{3j} - S(T)n_{4j} + N_{1\ j-1}(\sigma) \\
\frac{dn_{2j}}{d\sigma} &= G'_{1x}(X_0(T),T)n_{4j} + N_{2\ j-1}(\sigma) \\
\frac{dn_{3j}}{d\sigma} &= - B_2(T)n_{3j} - S_2(T)n_{4j} + N_{3\ j-1}(\sigma) \\
\frac{dn_{4j}}{d\sigma} &= B'_2(T)n_{4j} + N_{4\ j-1}(\sigma)
\end{aligned} \qquad (2.18)$$

where the exponentially decaying terms $N_{i\ j-1}(\sigma)$ are known successively. Equation (2.5d) dictates $n_{4j}(0) = - Q_j(T)$. Using lemma 1, $n_{3j}(0)$ can be selected such that $n_{3j}(\sigma)$ decays exponentially. It is evident then that $n_{ij}(\sigma)$ can be uniquely defined and decay exponentially as $\sigma \to \infty$. Equation (2.5c) then gives $P_{j+1}(T) = - n_{2j}(0)$.

Knowing the asymptotic expansions for the costate variables $p(t)$ and $q(t)$, the expansions for the optimal control $u(t)$ can easily be determined from (1.7). Since $m_{20}(\tau)$, $m_{40}(\tau)$, $n_{20}(\sigma)$ and $n_{40}(\sigma)$ are identically zero, it is seen explicitly that the optimal control does not have any zero-order boundary layers either at the initial point $t = 0$ or at the final point $t=T$.

3. ASYMPTOTIC VALIDITY

Let us define the partial sums

$$\begin{aligned}
(x)_N &= X_0(t) + \sum_{j=1}^{N} (X_j(t) + m_{1\ j-1}(t/\lambda) + n_{1\ j-1}(\tfrac{T-t}{\lambda}))\lambda^j \\
(p)_N &= P_0(t) + \sum_{j=1}^{N} (P_j(t) + m_{2\ j-1}(t/\lambda) + n_{2\ j-1}(\tfrac{T-t}{\lambda}))\lambda^j \\
(z)_N &= \sum_{j=0}^{N} (Z_j(t) + m_{3j}(t/\lambda) + n_{3j}(\tfrac{T-t}{\lambda}))\lambda^j \\
(q)_N &= \sum_{j=0}^{N} (Q_j(t) + m_{4j}(t/\lambda) + n_{4j}\ (\tfrac{T-t}{\lambda}))\lambda^j .
\end{aligned} \qquad (3.1)$$

Theorem:

Under the hypotheses H_1 and H_2, the two point boundary value problem (1.8) and (1.9) has a unique solution for λ sufficiently small such that for every integer $N \geq 0$,

$$x = (x)_N + \lambda^N x^*$$

$$p = (p)_N + \lambda^N p^*$$

$$z = (z)_N + \lambda^N z^*$$

$$q = (q)_N + \lambda^N q^*$$

where x^*, p^*, z^*, and q^* tend to zero as λ tends to zero uniformly for $0 \leq t \leq T$.

The limitations of available space preclude including the proof.

REFERENCES

1. L. Flatto and N. Levinson, "Periodic Solutions of Singularly Perturbed Systems," J. Rat. Mech. Anal., Vol. 4, pp. 943-950, 1955.
2. F. R. Gantmacher, The Theory of Matrices, Chelsea Publishing Company, New York, 1959.
3. C. Hadlock, "Singular Perturbations of a Class of Two Point Boundary Value Problems Arising in Optimal Control," Ph.D. Thesis, Dept. of Mathematics, Un. of Illinois, June 1970, CSL Report No. 481.
4. C. Hadlock, "Singular Perturbation of a Class of Boundary Value Problems Arising in Optimum Control," 8th Allerton Conf. on Circuit and System Theory, October 7-9, 1970.
5. F. Hoppensteadt, "Properties of Solutions of Ordinary Differential Equations with Small Parameters," to appear in Communications of Pure and Applied Mathematics.
6. P. Kokotović and P. Sannuti, "Singular Perturbation Method for Reducing the Model Order in Optimal Control Design," IEEE Transactions on Automatic Control, Vol. 13, pp. 377-384, August 1968.
7. R. E. Kalman, "Contributions to the Theory of Optimal Control," Bol.Soc. Math., Mexicana, pp. 102-119, 1960.
8. J. J. Levin, "Singular Perturbations of Nonlinear Systems of Differential Equations Related to Conditional Stability," Duke Math.J., Vol. 23, pp. 609-620, 1956.
9. R. E. O'Malley, Jr., "The Singularly Perturbed Linear State Regulator Problem."
10. P. Sannuti, "Singular Perturbation Method in the Theory of Optimal Control," Ph.D. Thesis, Dept. of Elec. Eng'g, Un. of Illinois, June 1968, CSL Report No. 379.
11. P. Sannuti and P. Kokotović, "Singular Perturbation Method for Near Optimum Design of High-Order Nonlinear Systems," Automatica, Vol. 5, pp. 773-779, December 1969.
12. V. A. Tupčiev, "Asymptotic Behavior of the Solution of a Boundary Problem for Systems of Differential Equations of First Order with a Small Parameter in the Derivative," Sov. Math. (Doklady), Vol. 3, pp. 302-305, 1962.
13. A. B. Vasil'eva, "Asymptotic Behavior of Solutions to Certain Problems Involving Nonlinear Differential Equations Containg a Small Parameter Multiplying the Highest Derivatives," Russian Mathematical Surveys, Vol. 18, No. 3, pp. 13-81, 1963.

AN ALGORITHM FOR EQUIVALENT REDUCED STATE DESCRIPTIONS FOR LUMPED LINEAR TIME INVARIANT SYSTEMS

K. S. Bajaj
Department of Electrical Engineering and Systems Science
Michigan State University

Abstract

A new method is presented for equivalent reduced state descriptions for Systems which are: 1. observable but not controllable, 2. controllable but not observable, 3. neither controllable nor observable.

Introduction

This paper presents a new method for equivalent reduced state descriptions for systems which are 1. observable but not controllable, 2. controllable but not observable, 3. neither controllable nor observable.

Algorithm

Before we can give the algorithm, we have to state some definitions.

Defin. 1. [1] A continuous time system $\dot{x} = Ax + Bu$ where $x = n$ vector, $u = r$ vector, $A = n \times n$ matrix, $B = n \times r$ matrix is completely state controllable $\Leftrightarrow$ the composite $n \times nr$ matrix $P = [B:AB:A^2B: \ldots (A)^{n-1}B]$ is of rank n.

Defin. 2. [1] A continuous time system $\dot{x} = Ax$, $y = Cx$ where $x = n$ vector, $y = m$ vector, $A = n \times n$ matrix, $c = m \times n$ matrix is completely observable $\Leftrightarrow$ the composite $n \times mn$ matrix $P = [C^*:A^*C^*: \ldots (A^*)^{n-1}C^*]$ is of rank n.

Defin. 3. [2] State Equivalence

State a_i is equivalent to a_j $\Leftrightarrow$ for all inputs u, the response segment of $\mathcal{A}$ (abstract oriented object) starting in state a_i is identical with the response segment of $\mathcal{A}$ starting in state a_j

$$[a_i = a_j] \Leftrightarrow [\overline{A}(a_i, u) = \overline{A}(a_j, u)] \; \forall u.$$

Theorem 1. Let the state equations be given in the form

$$\dot{x} = Ax + Bu$$
$$y = Cx$$

where
x = n vector
u = scalar
A = n X n matrix
B = n X 1 matrix
y = m vector
C = m X n

Assume that the system is not controllable, but observable then the equivalent reduced state description given by

$$\begin{aligned} \dot{\lambda} &= \hat{A}\lambda + \hat{B}u \\ Y &= \hat{C}\lambda \end{aligned} \qquad (1)$$

is obtained by the following procedure:

i. Compute the controllability matrix (n X n)

$$P = [B:AB:A^2B: \ldots A^{n-1}B]$$

ii. If the rank of P is n_c then, by row operations on P, one can produce a matrix of the form $\left[\frac{0}{\hat{P}}\right]$ where rank of $\hat{P}$ = # rows of $\hat{P} = n_c$. Call $\hat{P}$ the reduced controllability matrix.

iii. Now define $\hat{B}$ to be first column of the reduced P matrix and define

$$\hat{A} = [\hat{A}\hat{B}:\hat{A}^2\hat{B}: \ldots (\hat{A})^{n_c}\hat{B}][\hat{B}:\hat{A}\hat{B}: \ldots (\hat{A})^{n_c-1}\hat{B}]^{-1} \qquad (2)$$

where $[\hat{B}:\hat{A}\hat{B}:\hat{A}^2\hat{B}: \ldots (\hat{A})^{n_c-1}\hat{B}]$

is the nonsingular matrix formed from the first n_c columns of $\hat{P}$.

iv. $\hat{C}$ is found from

$$\begin{aligned} \hat{C}\hat{B} &= CB \\ \hat{C}\hat{A}\hat{B} &= CAB \\ &\vdots \\ \hat{C}(\hat{A})^{n_c-1}\hat{B} &= C(A)^{n_c-1}B \end{aligned}$$

or $$\hat{C} = [CB:CAB: \ldots CA^{n_c-1}B][\hat{B}:\hat{A}\hat{B}: \ldots (\hat{A})^{n_c-1}\hat{B}]^{-1} \qquad (3)$$

Proof.

i. Consider the matrix

$$P = \{\hat{B}:\hat{A}\hat{B}:\hat{A}^2\hat{B}: \ldots \hat{A}^{n-1}\hat{B}\}$$

if this matrix has rank n_c then it is easy to see that the matrix

$$\{\hat{B}:\hat{A}\hat{B}:\hat{A}^2\hat{B}:\dots(\hat{A})^{n_c-1}\hat{B}\}$$

formed from the first n_c columns of P is of rank n_c. Hence the system

$$\dot{\lambda} = \hat{A}\lambda + \hat{B}u$$

is controllable.

ii. Proof of the state equivalence of two systems. For

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

$$x(t) = e^{At}\{x(0) + \int_0^t e^{-A\tau} Bu(\tau)\, d\tau\}$$

$$y(t) = Ce^{At}\{x(0) + \int_0^t e^{-A\tau} Bu(\tau)\, d\tau\} \qquad (4)$$

Zero state output is given by

$$y_1^0(t) = Ce^{At}\int_0^t e^{-A\tau} Bu(\tau)\, d\tau$$

$$y_1^0(t) = C\int_0^t e^{A\tau} Bu(t-\tau)\, d\tau$$

we can express

$$e^{A\tau} = \sum_{i=0}^{n-1} \alpha_i(\tau) A^i$$

n: degree of minimal polynomial of A.

$$y_1^0(t) = C \sum_{i=0}^{n-1} \sum_{j=1}^{r} \gamma_{ij} A^i B_j$$

where

$$\gamma_{ij} = \int_0^t \alpha_i(\tau)\, U_j(t-\tau)\, d\tau = \text{scalar}$$

$$y_1^0(t) = C[B:AB:\dots A^{n-1}B][\underset{\sim}{\gamma}_{ij}] \qquad (5)$$

By performing row operations on

$$[B:AB:\dots A^{n-1}B]$$

we have P matrix as

$$[LB:LAB:LA^2B:\dots LA^{n-1}B] \qquad (6)$$

it is of the form

$$C = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline K & M \end{array}\right]$$

where K is $n_c \times n_c$ nonsingular matrix or we can write (6) as

$$[LB:LAL^{-1}\,LB:(LAL^{-1})^2\,LB:\dots(LAL^{-1})^{n-1}\,LB] \qquad (7)$$

where

$$LB = \left[\begin{array}{c} 0 \\ \hdashline \hat{B} \end{array}\right] \quad \text{and} \quad LAL^{-1} = \left[\begin{array}{c:c} P & 0 \\ \hdashline 0 & \hat{A} \end{array}\right]$$

Now we have

$$|A - \lambda I| = |P - \lambda I| \cdot |\hat{A} - \lambda I|$$

which shows that $\hat{A}$ satisfies the characteristic polynomial of A and they have n_c common eigenvalues. If we expand $e^{A\tau}$ and $e^{\hat{A}\tau}$ only in terms of A^n and $(\hat{A})^n$ for $n = 0, 1, 2, \dots, (n_c - 1)$ then χ_{ij} will be same in eqn. (5). From (5) we have

$$T y_1^0(t) = C[LB:LAB:\dots LA^{n_c-1}B][\chi_{ij}]$$
$$y_1^0(t) = T^{-1}C[\hat{B}:\hat{A}\hat{B}:\dots(\hat{A})^{n_c-1}\hat{B}][\chi_{ij}]$$
$$y_1^0(t) = \hat{C}[\hat{B}:\hat{A}\hat{B}:\dots(\hat{A})^{n_c-1}\hat{B}][\chi_{ij}] \qquad (8)$$

Comparing (5) and (8) we have

$$\hat{C}\hat{B} = CB$$
$$\hat{C}\hat{A}\hat{B} = CAB$$
$$\vdots$$
$$\hat{C}(\hat{A})^{n_c-1}\hat{B} = CA^{n_c-1}B$$

Hence the proof of the theorem.

Extension of the above theorem for cases where we have

u = r vector and

B = n X r matrix

is quite simple. After row reduction of the controllability matrix. First r columns of the row reduced matrix comprise of $\hat{B}$. While for $\hat{A}$ we have to pick n_c independent columns out of nr columns of controllability matrix and $\hat{A}$ can be found

$$[\hat{A}\hat{B}_1:(\hat{A})^2\hat{B}_1:\dots(\hat{A})^{n_c}\hat{B}_1][\hat{B}_1:\hat{A}\hat{B}_1:\dots(\hat{A})^{n_c-1}\hat{B}_1]^{-1} \qquad (9)$$

where $\hat{B}_1$ is the first column of $\hat{B}$ and $[\hat{B}_1:\hat{A}\hat{B}_1:\dots(\hat{A})^{n_c-1}\hat{B}_1]$ are the independent columns of row reduced controllability matrix $\hat{C}$.

We can pick any of the columns of $\hat{B}$ and find $\hat{A}$ or in general we can write

$$\hat{A} = [\hat{A}\hat{B}_{\gamma_1} : (\hat{A})^2 \hat{B}_{\gamma_1} : \dots (\hat{A})^{n_c} \hat{B}_{\gamma_1}][\hat{B}_{\gamma_1} : (\hat{A}) \hat{B}_{\gamma_1} : \dots (\hat{A})^{n_c - 1} \hat{B}_{\gamma_1}]^{-1} \quad (10)$$

Next we can find $\hat{C}$ by Eq. (3).

Example 1. Consider the system

$$\Sigma_1 : \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -6 \\ 1 & 0 & -11 \\ 0 & 1 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 4 \\ 5 \\ 1 \end{bmatrix} u$$

$$y = [1 \quad 2 \quad 3] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

which is not controllable but is observable. The controllability matrix is

$$\begin{bmatrix} 4 & -6 & 6 \\ 5 & -7 & 5 \\ 1 & -1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 0 \\ 4 & -6 & 6 \\ 1 & -1 & -1 \end{bmatrix}$$

$$\hat{B} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

$$\hat{A} = \begin{bmatrix} -6 & 6 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 4 & -6 \\ 1 & -1 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & -6 \\ 1 & -5 \end{bmatrix}$$

The observability matrix is

$$\begin{bmatrix} 1 & 1 & 3 \\ 2 & 2 & -46 \\ 3 & -46 & 231 \end{bmatrix}$$ and the system is observable.

In order to calculate $\hat{C}$ we use

$$CB = \hat{C}\hat{B}$$

$$CAB = \hat{C}\hat{A}\hat{B}$$

$$\hat{C} = [3 \quad 5]$$

Hence, we have the equivalent reduced state description

$$\Sigma_2: \begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \end{bmatrix} = \begin{bmatrix} 0 & -6 \\ 1 & -5 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} + \begin{bmatrix} 4 \\ 1 \end{bmatrix} u$$

$$y = \begin{bmatrix} 3 & 5 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}$$

On comparison, one sees that the transfer functions of Σ_1 and Σ_2 are identical.

Theorem 2. Let the state equations be given as

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

$$y = Cx$$

where u = r vector, A = n X n matrix

y = scalar, B = n X r matrix

C = 1 X n matrix

Assume the system is controllable but not observable. Then the reduced state description

$$\dot{\lambda} = \hat{A}\lambda + \hat{B}u$$

$$y = \hat{C}\lambda$$

is given by the following procedure:

i. Find the observability matrix $\underset{\sim}{0}$

$$[C^*:A^*C^*: \ldots (A^*)^{n-1}C^*]$$

ii. Find the reduced observability matrix $\hat{\underset{\sim}{0}}$.

iii. Then first column of $\hat{\underset{\sim}{0}}$ comprises of $(\hat{C})^*$ and

$$(\hat{A})^* = [(\hat{A})^*\hat{C}^*:(\hat{A}^*)^2\hat{C}^*: \ldots ((\hat{A})^*)^{n_c}\hat{C}^*][\hat{C}^*:\hat{A}^*\hat{C}^*: \ldots (\hat{A}^*)^{n_c-1}\hat{C}^*]^{-1} \tag{11}$$

iv. $\hat{B}$ is found from

$$\hat{C}\hat{B} = CB$$

$$\hat{C}\hat{A}\hat{B} = CAB$$

$$\vdots$$

$$\hat{C}(\hat{A})^{n_c-1}\hat{B} = C(A)^{n_c-1}B$$

or $$\hat{B}^* = [B^*C^*:B^*A^*C^*: \ldots B^*(A^*)^{n_c-1}C^*][\hat{C}^*:\hat{A}^*\hat{C}^*: \ldots (\hat{A}^*)^{n_c-1}\hat{C}^*]^{-1} \tag{12}$$

Now extension of the above theorem for multi-output system is simple. Let C = m X n matrix

y = m vector

then after row reduction of $\underset{\sim}{0}$ which is (n X mn) matrix first m columns of $\underset{\sim}{\hat{0}}$ comprise of $\hat{C}^*$ and

$$(\hat{A})^* = [\hat{A}^*\hat{C}^*_{\gamma_1} : (\hat{A}^*)^2\hat{C}^*_{\gamma_1} : \dots (\hat{A}^*)^{n_c}\hat{C}^*_{\gamma_1}][\hat{C}^*_{\gamma_1} : (\hat{A}^*)\hat{C}^*_{\gamma_1} : \dots (\hat{A}^*)^{n_c-1}\hat{C}^*_{\gamma_1}]^{-1} \tag{13}$$

where

$$[\hat{C}^*_{\gamma_1} : \hat{A}^*\hat{C}^*_{\gamma_1} : \dots (\hat{A}^*)^{n_c-1}\hat{C}^*_{\gamma_1}]$$

are the independent columns of reduced $\underset{\sim}{\hat{0}}$.

$\hat{B}^*$ can be found from equation (12).

Theorem 3. Let the state equations be given as

$\dot{x} = Ax + Bu$

$y = Cx$

where x = n vector

u = r vector

A = n X n matrix

B = n X r matrix

y = m vector

C = m X n matrix

and the system is neither controllable nor observable then the equivalent reduced system is given by the following procedure:

i. Find the rank of controllability matrix, call it nc_1.

ii. Find the rank of observability matrix, call it nc_2.

iii. If $nc_1 < nc_2$ then apply theorem 1 to get reduced state description.

iv. If $nc_2 < nc_1$ then apply theorem 2 to get reduced state description.

v. If $nc_2 = nc_1$ then either of the theorems 1 or 2 can be used.

Proof. For $nc_1 < nc_2$ if we use theorem (1) we get the minimum possible dimension of $\hat{A}$ matrix from the reduced controllability matrix and $\hat{C}$ is being obtained from $\hat{C}\hat{A}^{\nu}\hat{B} = CA^{\nu}B$ for minimum possible ν. Hence we get the reduced state description in one stroke.

Now if we use theorem (2) for this case, then once we get $\hat{C}'$, $\hat{B}'$, $\hat{A}'$ we have to check the system

$$\begin{aligned}\dot{Z} &= \hat{A}'Z + \hat{B}'u \\ y' &= \hat{C}'Z\end{aligned}$$

again, for controllability this time and reduce it again.

Example 2. Consider the system

$$\Sigma_1: \begin{bmatrix}\dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & -1\end{bmatrix}\begin{bmatrix}x_1 \\ x_2 \\ x_3\end{bmatrix} + \begin{bmatrix}5 \\ 3 \\ 1\end{bmatrix} u$$

$$y = [1 \quad 1 \quad 1]\begin{bmatrix}x_1 \\ x_2 \\ x_3\end{bmatrix}$$

which is neither controllable nor observable.

1. Controllability matrix

$$C = \begin{bmatrix}5 & 5 & 5 \\ 3 & 3 & 3 \\ 1 & 1 & 1\end{bmatrix} \qquad \text{is of rank } 1$$

2. Observability matrix

$$\underset{\sim}{0} = \begin{bmatrix}1 & 2 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & 1\end{bmatrix} \qquad \text{is of rank } 2$$

According to theorem 3 we will apply theorem 1 to get reduced state description for the system.

From reduced controllability matrix

$$\hat{A} = 1$$

$$\hat{B} = 1$$

From $$CB = \hat{C}\hat{B}$$

$$\hat{C} = 9$$

Hence reduced state description is given by

$$\Sigma_2: \begin{aligned}\dot{\lambda} &= \lambda + u \\ y &= 9\lambda\end{aligned}$$

Acknowledgment.

The author is thankful to Dr. J. A. Resh and Dr. J. S. Frame for many useful discussions.

References

1. Ogata, K. State Space Analysis of Control Systems. Prentice-Hall, Inc., 1967.
2. Zadeh, L. A. and C. A. Desoer. Linear System Theory, The State Space Approach. McGraw-Hill Book Co., Inc., New York, New York, 1963.
3. Kwakernaak, H., "On the Reduction of the System x = Ax + Bu, y = Cx to its minimal equivalent," IEEE Trans. on Circuit Theory, Dec. 1963.
4. Kalman, R. E., "Mathematical Description of Linear Dynamical Systems," J. SIAM Control, 1 (1963), pp. 152-192.

A NEW FILTER FOR OPTIMAL TRACKING IN DENSE MULTITARGET ENVIRONMENTS

R.A. SINGER and R.G. SEA
Hughes Aircraft Company
Ground Systems Group
Fullerton, California

ABSTRACT

In dense multitarget environments, a tracking filter is occasionally updated with sensor returns not originating from the vehicle being tracked. As a result the traditional optimal filters no longer perform optimally and the calculated error covariance matrix differs substantially from the actual tracking error behavior. In this paper a new tracking filter is derived that is optimal under these conditions, and analytical means are developed for performance assessment of any arbitrary tracking filter.

I. INTRODUCTION

When "optimal" tracking filters such as the Kalman filter are used in surveillance systems, it frequently happens that when the target density is high, the actual and theoretical tracking performance statistics differ substantially, even for those regimes for which linear equations are valid. The cause of this deviation is the neglect of the effects of data correlation uncertainties by traditional tracking filter analyses. Inherent in the formulation of such optimal filters as the Kalman filter, for example, is the assumption that the sensor return used for track update originates from the vehicle being tracked. In high density multitarget environments, however, it is not generally possible to always select (i.e., correlate) the correct return (if it exists) from among potentially many candidates. Hence an incorrect return, namely, one not originating from the tracked target, will sometimes be chosen and used by the tracking filter. Since the tracking filter measurement model is violated by such a choice, the theoretical tracking covariance matrix becomes overly optimistic.

In this paper a new optimal tracking filter is derived that properly accounts for, and minimizes, the effects of correlation errors in surveillance systems. Both the filter gain matrix and the filter error covariance matrix of this new filter are shown to depend on the correlation statistics. When target density decreases, the new filter reduces in the limit to the traditional Kalman filter.

Also developed in this paper is a means for quantitatively assessing the performance of any standard tracking filter in a correlation error environment. This contribution will permit the tracking filter designer to conduct a more realistic a priori tracking performance analysis suitable for surveillance system error allocation, will help indicate those regions in which the on-line tracking statistics provided by an arbitrary filter cannot be used with confidence by the surveillance system, will provide realistic estimates of the corrections required in these regions, and will permit an analytical basis for tracking filter selection.

That an "optimal" filter will occasionally utilize incorrect measurement data was recognized in [1]. In that paper, however, the form of the measurment equations corresponding to an incorrect measurement differs significantly from that which arises in the surveillance system tracking applications of interest in the present paper. Also, in Nahi's paper, the occurrence of an incorrect measurement is independent of any of the estimation errors or measurement noise samples, whereas in tracking problems the occurrence of an incorrect measurement depends on the estimation errors, measurement noise samples, distributions of competing sensor returns and the decision rule used to select the return (i.e., measurement) to be used for updating the track. This dependence introduces correlation between certain random variables which are not normally correlated in traditional estimation problems.

A preliminary development of a new, reoptimized filter for realistic surveillance system tracking applications is described in [2]. In that paper, however, the filter equations were not developed in the generality appearing here, and the performance analysis of the standard Kalman filter was based on the generally unrealistic simplifying assumption that when such a filter is implemented, optimal correlation performance will be obtained. In this paper, these limitations are removed and in addition a more reliable means of determining the performance of any standard filter in dense environments is developed. A realistic tracking example is evaluated parametrically, permitting relative comparisons of the new and standard tracking filters.

II. MATHEMATICAL BACKGROUND

A. Dynamic Equations of Motion

The discretized equations of motion for the object being tracked may be modeled by $x(k+1) = \phi(k+1,k)\,x(k) + G(k)u(k)$ (1)
where $x(k)$ is an nx1 state vector evaluated at time kT, ϕ is the nxn transition matrix obtained from the original continuous time dynamical equations of vehicle motion, $G(k)$ is an nxp input distribution matrix, and $u(k)$ is a px1 state excitation vector. The observation equation representing sensor measurements of the object being tracked may be assumed to satisfy $y(k) = H(k)x(k) + v(k)$ (2)
where $y(k)$ is the Nx1 sensor measurement vector, $H(k)$ is the observation matrix, and $v(k)$ is white observation noise with covariance $R(k)$.

B. Classical Filter Formulation

The general form of the tracking filter used for track updating is

$$\hat{x}(k/k) = \hat{x}(k/k-1) + K(k)\left[y(k) - H\hat{x}(k/k-1)\right] \tag{3}$$

$$\hat{x}(k+1/k) = \phi(k+1,k)\,\hat{x}(k/k) \tag{4}$$

where $\hat{x}(i/j)$ is the filter estimate of $x(i)$ given measurements through time j, and where $K(k)$ is the filter gain, or correction, matrix at time k.

Many candidate filter types may be considered (e.g., [3-7]), each differing primarily in the calculation of the gain matrix $K(k)$. Attention will be given later in this paper to the classical Kalman filter, which satisfies [3]:

$$K(k) = P(k/k-1)H'(k)\left[H(k)P(k/k-1)H'(k) + R(k)\right]^{-1} \tag{5}$$

In this equation, $P(i/j)$ is the error covariance of the estimate $\hat{x}(i/j)$, satisfying

$$P(k/k) = P(k/k-1) - P(k/k-1)H'(k)\left[H(k)P(k/k-1)H'(k) + R(k)\right]^{-1}H(k)P(k/k-1) \tag{6}$$

$$P(k+1/k) = \phi(k+1,k)P(k/k)\phi'(k+1,k) + G(k)Q(k)G'(k) \tag{7}$$

where $Q(k)$ is the covariance of the state excitation noise vector $u(k)$, assumed to be white.

C. The Data Correlation Algorithm

An efficient algorithm that is often implemented for correlating sensor returns with stored tracks involves the use of gates and a "statistical distance" test. This correlation technique is optimized and described in detail in [8], where correlation performance and relative computational requirements are presented parametrically as a function of gate size, target and false alarm densities, tracking and measurement accuracies, and detection probability. In this algorithm, a gate, i.e., a region in the measurement space, is formed for each stored track, the center of the gate being located at the predicted measurement $\hat{y}(k/k-1) = H\hat{x}(k/k-1)$ for the track. Specifically, the gate $\tilde{G}$ is defined as

$$\tilde{G} = \left\{ \tilde{y} : |\tilde{y}_i| \le K_i \sigma_i ,\ i=1, \ldots, N \right\} \tag{8}$$

where $\tilde{y} = y - \hat{y}(k/k-1)$ is the residual associated with a return y, $\tilde{y}_i$ is the i^{th} component of the vector $\tilde{y}$, K_i is the normalized gate size (a dimensionless parameter) in the i^{th} measurement coordinate, and σ_i is the standard deviation of the i^{th} component $(\tilde{y}_i)$ of the residual.

The normalized gate widths K_i should be large enough to ensure a high probability that the correct sensor return (if detected) will fall within the gate. If no sensor returns fall in the gate, the track will not be updated and missing data is declared. If exactly one return falls in the gate, the track will be updated with that return. If more than one return falls in the gate, the "statistical distance" of each return from the center of the gate is computed and compared, and the return with the smallest "distance" is declared to be correlated with the track. The square of the norm of the distance associated with a particular return y is defined as

$$d^2(y) = (y - \hat{y}(k/k-1))' V^{-1}(k)\,(y - \hat{y}(k/k-1)) \tag{9}$$

where V(k) is the residual covariance matrix, which can be expressed in terms of the tracking filter parameters as $V(k) = H(k)P(k/k-1)H'(k) + R(k)$ (10)
The purpose of the gate is to limit the number of returns that must be compared in this fine correlation test, thereby limiting computational requirements.

D. Performance of the Correlation Algorithm When Tracking Statistics are Known Correctly

The performance of the correlation algorithm is characterized by two parameters: P_G, the probability that a correlation will be made, and P_C, the probability that a correlation is correct, given that a correlation is made. General expressions for these parameters were derived in [8]. The particular forms of these expressions for the common case where the number of measurement variables is two (e.g., range and azimuth) are:

$$P_G = 1 - e^{-n_{TF}K_1K_2}\left[1 - \frac{P_D}{2\pi\sqrt{|V|}}\int_{-K_1\sigma_1}^{K_1\sigma_1}\int_{-K_2\sigma_2}^{K_2\sigma_2} e^{-\frac{1}{2}\tilde{y}' V^{-1}\tilde{y}}\,d\tilde{y}\right] \tag{11}$$

$$P_C = \frac{P_D}{P_G}\,\frac{1}{2\pi\sqrt{|V|}}\int_{-K_1\sigma_1}^{K_1\sigma_1}\int_{-K_2\sigma_2}^{K_2\sigma_2} e^{-\frac{1}{2}\tilde{y}' V_1^{-1}\tilde{y}}\,d\tilde{y} \tag{12}$$

where

$$V_1 = \frac{1}{1 + \frac{\pi}{2}n_{TF}\sqrt{1-\rho^2}}\,V, \quad V = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{13}$$

n_{TF} is the expected number of sensor returns (due to targets, false alarms, etc., but excluding the correct return) falling within a region of ± one standard deviation of distance, in each measurement coordinate, from the predicted measurement, V is the covariance matrix of the residual, defined in (10), and P_D is the sensor's probability of detection for the particular track in question.

The performance of large gates, which are frequently implemented in tracking systems, may be approximated by the following infinite gate ($K_1, K_2 = \infty$) expressions:

$$P_{G_\infty} = 1, \quad P_{C_\infty} = \frac{P_D}{1 + \frac{\pi}{2}n_{TF}\sqrt{1-\rho^2}} \tag{14}$$

E. Correlation Performance When Tracking Statistics are Imprecisely Known

Often the tracking error statistics calculated by a filter are not equal to the actual tracking error statistics. This occurs, for example, when suboptimal filters such as the α-β filter, are used. Even when the Kalman filter is used and the dynamics modelling is valid, the tracking statistics provided by the filter covariance equations may be incorrect because correlation errors are neglected.

Since the correlation algorithm discussed earlier was developed using the statistics of the actual residual covariance matrix V, both its implementation (in terms of gate size and orientation) and performance will, in general, change when the residual covariance V, used in the correlation algorithm, is calculated by an on-line filter in which computed tracking statistics are incorrect. For example, if the true (actual) covariance is larger than the covariance calculated by the filter, the gate size will generally be smaller than desired, resulting in degraded correlation performance. Furthermore, even if the gates are sufficiently large so that sensitivity of performance to changes in gate size are small, the statistical distance test will normalize the residuals of all returns in the gate with respect to an incorrect residual covariance matrix, leading to suboptimal performance of this test.

Correlation performance in these circumstances depends jointly on the actual residual covariance V and the residual covariance $\hat{V}$ estimated by the tracking filter. In the following, the notation ∧ above a symbol refers to the value of this quantity as estimated by the tracking filter; the absence of this symbol means the actual value of the quantity. For example $\hat{n}_{TF}$ is the target density normalized to $\hat{V}$ (i.e., $\hat{\sigma}_1$, $\hat{\sigma}_2$, ...) rather than to V ($\sigma_1, \sigma_2, \ldots$)

The probability of correct correlation may be calculated as follows: When the residual $\tilde{y}$ for the proper return, assumed detected, lies within the gate set up by the filter, a correct correlation will occur when no other returns have residuals in the region $D(\tilde{y}) \cap \tilde{G}$, where

$$D(\tilde{y}) = \left\{ \tilde{z} : \tilde{z}'\hat{V}^{-1}\tilde{z} < \tilde{y}'\hat{V}^{-1}\tilde{y} \right\} \tag{15}$$

Unless $\hat{V} = \alpha V$, where α is any scalar, the return selected to be "closest" by such a test may not be the most likely return for the track. Following an analogous approach to that in [8], the conditional probability of correlation is determined to be

$$P(\text{correct} \,|\, \text{det}, \tilde{y}) = \exp\left[\frac{-\hat{n}_{TF} C_N \pi \sqrt{|\hat{V}|}\,(\tilde{y}'\hat{V}^{-1}\tilde{y})^{N/2}}{2\hat{\sigma}_1 \ldots 2\hat{\sigma}_N}\right] \tag{16}$$

where C_N is a constant ($C_1 = 2/\pi$, $C_2 = 1$, $C_3 = 4/3$, $C_4 = \pi/2$, ...)
To eliminate the conditioning on the correct residual $\tilde{y}$, it is necessary to use the probability density function for $\tilde{y}$ and integrate with respect to $\tilde{y}$ over the gate:

$$P(\text{correct/det}) = \int_{\tilde{G}} \exp\left[\frac{-\hat{n}_{TF} C_N \pi \sqrt{|\hat{V}|}\,(\tilde{y}'\hat{V}^{-1}\tilde{y})^{N/2}}{2\hat{\sigma}_1 \ldots 2\hat{\sigma}_N}\right] p(\tilde{y})\,d\tilde{y} \tag{17}$$

$$\text{where } p(\tilde{y}) = \frac{1}{(2\pi)^{N/2}\sqrt{|V|}}\, e^{-\frac{1}{2}\tilde{y}'V^{-1}\tilde{y}} \tag{18}$$

It should be noted that the p.d.f. for $\tilde{y}$ depends on the actual covariance matrix V, not on the assumed covariance matrix $\hat{V}$.

Finally, P_C may be expressed by $P_C = P(\text{correct/det})P_D/P_G$. (19)

In the case of a two dimensional measurement (N=2) this becomes, after some manipulation,

$$P_C = \frac{P_D}{P_G} \frac{1}{\sqrt{\left| I + \frac{\pi}{2}\hat{n}_{TF}\sqrt{1-\hat{\rho}^2}\, V\hat{V}^{-1} \right|}} \left\{ \frac{1}{2\pi\sqrt{|V_1|}} \int_{-K_1\hat{\sigma}_1}^{K_1\hat{\sigma}_1} \int_{-K_2\hat{\sigma}_2}^{K_2\hat{\sigma}_2} e^{-\frac{1}{2}\tilde{y}'V_1^{-1}\tilde{y}}\, d\tilde{y} \right\} \tag{20}$$

where

$$V_1 = \left[I + \frac{\pi}{2}\hat{n}_{TF}\sqrt{1-\hat{\rho}^2}\, V\hat{V}^{-1} \right]^{-1} V, \quad \hat{V} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 \\ \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 \end{bmatrix} \tag{21}$$

The probability P_G that a correlation will be made is found similarly to be

$$P_G = 1 - e^{-\hat{n}_{TF} K_1 K_2 \left[1 - \frac{P_D}{2\pi \sqrt{|V|}} \int_{-K_1\hat{\sigma}_1}^{K_1\hat{\sigma}_1} \int_{-K_2\hat{\sigma}_2}^{K_2\hat{\sigma}_2} e^{-\frac{1}{2}\tilde{y}' V^{-1} \tilde{y}} \, d\tilde{y} \right]} \tag{22}$$

When the gate sizes approach infinity, the above expressions reduce to

$$P_{G_\infty} = 1, \; P_{C_\infty} = \frac{P_D}{\sqrt{\left| I + \frac{\pi}{2} \hat{n}_{TF} \sqrt{1 - \hat{\rho}^2}\, V\hat{V}^{-1} \right|}} \tag{23}$$

When $\hat{V} = \alpha V$, where α is any scalar, (23) reduces to (14). It can be shown that, as expected, P_{C_∞} attains its maximum value when $\hat{V}$ satisfies this condition.

III. TRACKING PERFORMANCE IN DENSE ENVIRONMENTS

A. Performance of an Arbitrary Filter

When the sensor provides a new set of candidate measurements for the vehicle being tracked, three distinct situations can occur: (1) the track is not updated (the gate is empty), or (2) the track is updated with either (a) an incorrect return or (b) a correct return.

If no return appears in the gate, the filter gain vector K(k) is zero, so that

$$\hat{x}(k/k) = \hat{x}(k/k-1), \; P(k/k, \text{ no update}) = P(k/k-1, \text{ no update}) \tag{24}$$

where the term on the right is the a posteriori covariance using measurements through k-1 as well as the information that the track was not updated at k.

When at least one return falls within the gate, a positive correlation is declared, and the filter is updated by

$$\hat{x}(k/k) = \hat{x}(k/k-1) + K(k) \left[y^*(k) - H(k)\hat{x}(k/k-1) \right] \tag{25}$$

where K(k) is the filter gain matrix and y*(k) is the return selected for filter update.

When the positive correlation decision is incorrect (conditional probability $(1-p_C)$), the selected measurement y*(k) is not the proper measurement, and originates from a false alarm (e.g., thermal or process noise), from another vehicle in the airspace, or from clutter. Assuming that incorrect returns are uniformly distributed within the gate region, and since the gate is centered at $\hat{y}(k/k-1)$, it follows that the output equation when an incorrect correlation occurs is

$$y^*(k) = H(k)\hat{x}(k/k-1) + w(k) = \hat{y}(k/k-1) + w(k) \tag{26}$$

where w(k) has zero mean and covariance $R_W(k)$, where $R_W(k)$ is the covariance of the closest return to $\hat{y}(k/k-1)$, given that this closest return is an incorrect return and that the incorrect returns are uniformly distributed within the gate. A general expression for $R_W(k)$ is given in the appendix, together with the particular form applicable to the example discussed later. The resulting filtering error (where $e(k/j) = x(k) - \hat{x}(k/j)$ denotes the general estimation error) is $e(k/k) = e(k/k-1) - K(k)w(k)$ so that the corresponding covariance, assuming whiteness of w(k), is

$$P(k/k, \text{ incorrect update}) = P(k/k-1, \text{ incorrect update})$$

$$-B'(k,G)K'(k) - K(k)B(k,G) + K(k)R_W(k)\,K'(k) \tag{27}$$

$$\text{where } B(k,G) \triangleq E\left[w(k)e'(k/k-1)/\text{incorrect update} \right] \tag{28}$$

It is shown in the appendix that B(k, G) = 0 for the correlation algorithm under consideration in this paper.

When a correct correlation is made, given that a positive correlation occurs (conditional probability p_C), the selected measurement y* (k) is the proper measurement y(k), satisfying $y^*(k) = y(k) = H(k)x(k) + v(k)$ where v(k) is white noise with covariance R(k). The error in the updated estimate is thus

$$e(k/k) = e(k/k-1) - K(k)\left[H(k)e(k/k-1) + v(k) \right] = \left[I - K(k)H(k) \right] e(k/k-1) - K(k)v(k) \tag{29}$$

In obtaining the corresponding covariance equation, it must be observed that $e(k/k-1)$ and $v(k)$, while statistically independent in an unconditional sense, are not statistically independent when conditioned on the fact that a correct correlation is made. The reason for this is that when a correct correlation is made, the residual $He(k/k-1) + v(k)$ must lie in the gate and the normalized (with respect to $\hat{V}$) magnitude of this residual must be smaller than that of any other residual in the gate. The resulting updated covariance matrix is, therefore

$$\begin{aligned} P(k/k, \text{ correct update}) &= \left[I - K(k)H(k)\right] P(k/k-1, \text{ correct update}) \left[I - K(k)H(k)\right]' \\ &- \left[I - K(k)H(k)\right] A'(k,G)K'(k) - K(k)A(k,G) \left[I - K(k)H(k)\right]' + K(k)R_2(k,G)K'(k) \end{aligned} \tag{30}$$

where

$$\begin{aligned} A(k,G) &\triangleq E\left[v(k)\, e'(k/k-1)/\text{correct correlation}\right] \\ R_2(k,G) &\triangleq E\left[v(k)v'(k)/\text{correct correlation}\right] \end{aligned} \tag{31}$$

This differs from the standard Kalman filter equations in which $A(k,G) = 0$, $R_2(k,G) = R(k)$, and $P(k/k-1)$ is unconditional, since updates are not conditioned as they are here. Expressions for A, R_2, and the important conditional forms of $P(k/k-1)$ are given in the appendix.

From (27) and (30), it follows that the total tracking filter error covariance matrix, given a return within the correlation gate, and after averaging over the probabilities of correct and incorrect correlation, is therefore

$$\begin{aligned} P(k/k, \text{ update}) &= P(k/k, \text{ incorrect update})(1-P_c) + P(k/k, \text{ correct update})\, P_c \\ &= P(k/k-1, \text{ update}) - K(k) \left\{ P_c(k) \left[H(k)P_2(k,G) + A(k,G)\right] + \left[1-P_c(k)\right] B(k,G) \right\} \\ &- \left\{ P_c(k) \left[P_2(k,G)H'(k) + A'(k,G)\right] + \left[1-P_c(k)\right] B'(k,G) \right\} K'(k) + K(k)R^*(k)K'(k) \end{aligned} \tag{32}$$

where

$$\begin{aligned} P(k/k-1, \text{ update}) &= P(k/k-1, \text{ incorrect update})(1-P_c) + P(k/k-1, \text{ correct update})P_c \quad (33) \\ P_2(k,G) &\triangleq P(k/k-1, \text{ correct update}) \\ R^*(k) &= P_c(k) \left[H(k)P_2(k,G)H'(k) + R_2(k,G) + H(k)A'(k,G) + A(k,G)H'(k)\right] \\ &\quad + \left[1-P_c(k)\right] R_w(k) \end{aligned}$$

The expression $P_c(k)$ reflects that P_c changes from sample to sample.

Averaging over the probabilities of updating and not updating the track, it follows that the unconditional tracking filter error covariance matrix, when the arbitrary filter gains $K(k)$ are used, satisfies

$$\begin{aligned} P(k/k) &= P(k/k, \text{ update})P_G + P(k/k, \text{ no update})(1 - P_G) \\ &= P(k/k-1) - P_G K(k) \left\{ P_c(k) \left[H(k)P_2(k,G) + A(k,G)\right] \right. \\ &\quad \left. + \left[1-P_c(k)\right] B(k,G) \right\} - P_G(k) \left\{ P_c(k) \left[P_2(k,G)H'(k) + A'(k,G)\right] \right. \\ &\quad \left. + \left[1-P_c(k)\right] B'(k,G) \right\} K'(k) + P_G(k)\, K(k)\, R^*(k)\, K'(k) \end{aligned} \tag{34}$$

Expressions for $P(k/k-1, \text{ update})$ and $P(k/k-1, \text{ no update})$ are given in the appendix.

The actual performance of any tracking filter used in a dense multitarget environment can be estimated from (32) and (34). The typical filter determines its gain matrix $K(k)$ from its calculated covariance matrix $\hat{P}(k/k-1)$. Hence in evaluating its actual covariance matrix $P(k/k)$, both $P(k/k-1)$ and $\hat{P}(k/k-1)$ must be continually updated; $P(k/k)$ is a function of $K(k)$ and $P(k/k-1)$ as in (34), and $K(k)$ is a function of $\hat{P}(k/k-1)$, where $\hat{P}(k/k-1)$ is updated by the traditional tracking equations, which ignore correlation errors.

B. Formulation and Performance of the Optimal Filter

When a correlation is declared, and the tracking filter updates the track using the arbitrary gain K(k), the error covariance equation (32) results. Setting the first variation to zero as K(k) varies, ignoring second order terms, and recalling that P(k/k-1) is a given constant matrix, it follows that the optimal gain vector K*(k) satisfies

$$K^*(k) = \left\{ P_c(k) \left[P_2(k,G)H'(k)+A'(k,G) \right] + \left[1-P_c(k) \right] B'(k,G) \right\} R^{*^{-1}}(k) \qquad (35)$$

and that the corresponding optimal error covariance is given by

$$P^*(k/k,\ \text{update}) = P(k/k-1,\ \text{update}) - \left\{ P_c(k) \left[P_2(k,G)H'(k) + A'(k,G) \right] \right.$$

$$\left. + \left[1-P_c(k) \right] B'(k,G) \right\} R^{*^{-1}}(k) \left\{ P_c(k) \left[H(k)P_2(k,G)+A(k,G) \right] + \left[1-P_c(k) \right] B(k,G) \right\} \qquad (36)$$

Equations (35) and (36) give the tracking filter gain and the corresponding (minimized) error covariance matrix for the reoptimized tracker when correlation errors are considered. This new filter incorporates the correlation statistics directly into its gain and covariance matrix equations.

If, in addition to the assumed proper dynamics modelling, this reoptimized filter is initialized properly, then iterative application of (24) and (36) will assure that $\hat{V} = V$. In this case, the terms P_C, P_G, A, R_W, B, and R_2, and the conditional forms of P(k/k-1) depend only on V, and the optimal filter equations reduce to those considered in [2].

It is easily shown from (32), (35), and (36) that, for an arbitrary gain matrix K(k), and for the case when a return falls within the correlation gates:

$$P(k/k,\ \text{update}) = P^*(k/k,\ \text{update}) + \left[K(k) - K^*(k) \right] R^*(k) \left[K(k) - K^*(k) \right]'$$

Since the kernel R* (k) of the quadratic term is positive definite, the last expression is always at least positive semidefinite when K(k) ≠ K*(k). Hence the expression for K*(k) given in (35) is an optimal tracking filter gain matrix.

Averaging over the probabilities of updating and not updating the filter, it follows that the unconditional tracking filter error covariance matrix, when the optimal gains K* (k) are used, satisfies

$$P^*(k/k) = P(k/k-1) - P_G(k) \left\{ P_C(k) \left[P_2(k,G)H'(k) + A'(k,G) \right] + \left[1-P_C(k) \right] B'(k,G) \right\}$$

$$\cdot\ R^{*^{-1}}(k) \left\{ P_C(k) \left[H(k)P_2(k,G) + A(k,G) \right] + \left[1-P_c(k) \right] B(k,G) \right\} \qquad (37)$$

As the number of incorrect returns vanishes and as the gate region approaches infinity, so that $P_c \rightarrow 1$, $P_G \rightarrow P_d$, $A(k,G) \rightarrow 0$, $R_2(k,G) \rightarrow R(k)$, $P_2(k,G) \rightarrow P(k/k-1)$, $B(k,G) \rightarrow 0$, (37) reduces to

$$P(k/k) = P(k/k-1) - P_D(k)P(k/k-1)H'(k) \left[H(k)P(k/k-1)H'(k) + R(k) \right]^{-1} H(k)P(k/k-1)$$

As expected, this is the expression for the unconditional covariance provided by the standard Kalman filter.

IV. AN AIRCRAFT TRACKING ILLUSTRATION

The performances of the new reoptimized filter and the standard Kalman filter were evaluated parametrically for a second order, two dimensional aircraft tracking application in terms of relative accuracy and the reliability of tracking accuracy assessment provided by the Kalman filter. The aircraft, as in [5], follows essentially constant velocity trajectories except for occasional maneuver acceleration inputs modelled for tracking purposes as white noise. The tracking sensor provides the two dimensional "pseudo-measurement" vector consisting of x and y position. The dynamical equations of motion and the measurement equation are identical to those considered in [2], with the following exceptions. Values for the correlation coefficients ρ_Q and ρ_R between the x and y maneuver accelerations and between the x and y position measurements, respectively, were permitted to range between 0 and 0.9. Because the aircraft generally maneuvers either normal to (circular turn) or along (linear acceleration) its current trajectory, rather than independently along the tracking sensor's x and y reference axes, nonzero values for ρ_Q occur in almost all realistic situations. Because the tracking sensor provides range and angle information directly, the process of extracting x and y measurements results in nonzero values of ρ_R, the correlation coefficient between errors in x and y position "measurements." In [2], it was necessary to assume that ρ_Q and ρ_R were zero in order to obtain $\hat{V} = \alpha V$, (where α is some scalar), a necessary condition to be able to use the performance equations in [2], which are based on the inherent assumption of ideal correlation performance. Since the tracking and correlation performance equations developed in this paper permits handling the realistic situations of general V and $\hat{V}$, no such simplifications were required or utilized.

Tracking performance behavior can be characterized in terms of the dynamical parameters $T^2\sigma_M/\sigma_R$, ρ_R, and ρ_Q, and the environment parameters P_D, G, and β. In these expressions T is the sensor sample interval, $\sigma_M{}^2$ and $\sigma_R{}^2$ are the variances of the target maneuver and sensor measurement error, respectively, G is the gate size, which was taken as infinity in order to maximize correlation accuracy for the parameters in this example (see the curves in [8]), and β is the density of incorrect returns per unit area. These parameters were permitted to range over several orders of magnitude. For each case the correlation dependent terms were calculated in real time, using the relationships shown in the appendix and in (23). Tracking performance was obtained using the appropriate covariance analysis for each filter. Figure 1 illustrates the comparative results for one set of typical conditions existing at the one-hundredth iteration of each filter. The ratio of the normalized variance of filtered position error $(P_{11}(100)/\sigma_R{}^2)$ to the normalized variance of filtered position error provided by the standard Kalman filter when no correlation uncertainties exist, is plotted versus a normalized measure of false return density $(\beta\sigma_R{}^2)$ for $\rho_Q = 0$ and 0.9, for $\rho_R = 0$ and 0.9 and for $T^2\sigma_M/\sigma_R = 10$. Results for both the reoptimized filter and the standard Kalman filter are indicated. Since logarithms are used, it follows that when the ordinate is zero, the filter in question provides a performance identical to that of the Kalman filter in the absence of correlation errors.

The three generic regions of comparative filter performance discussed in [2] are again apparent in this figure. Also since target maneuver behavior dominates sensor measurement inaccuracy in this case, the role of ρ_Q is significant while that of ρ_R is minor. The opposite situation occurs when $T^2\sigma_M/\sigma_R = 0.1$. Since the effects of ρ_R and ρ_Q are important, the development of the general performance equations in this paper permit, for the first time, reliable tracking performance assessment for almost the entirety of target-sensor parameter values.

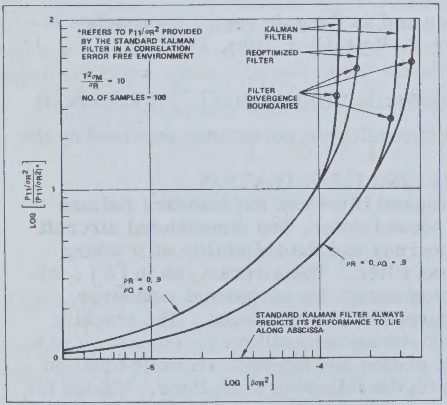


Figure 1. Comparative Tracking Filter Performance

V. CONCLUSIONS

The performance equations developed in this paper for an arbitrary tracking filter being utilized in a dense multitarget environment improves considerably the reliability of estimates that can be made of on line surveillance system performance, often removing the need for extensive Monte Carlo tracking simulation in the design phase of such systems. Furthermore the performance equations can be used, as in the example, to determine those operating regions in which a given tracking filter will perform unsatisfactorily either in terms of tracking accuracy directly or in terms of providing meaningful statistics on its performance.

The reoptimized filter as an alternative to various standard filters, such as the Kalman filter, for high density environment applications, offers two primary advantages. In the region for which target density effects first become important (this region may be determined directly from the covariance analysis discussed previously), the new filter and the standard Kalman filter perform similarly, albeit considerably worse than in a correlation error free environment. Only the new filter provides reliable estimates of its on-line performance in this region, a factor particularly critical for systems making weapon and launch decisions based on statistics including tracking accuracy. In higher density environments, the "reoptimized" filter performs significantly better than the standard Kalman filter, resulting frequently in much improved tracking accuracy and in much longer track lives. The cost for this benefit is, of course, considerably greater computational burden unless stored (precalculated) gain versions of this filter are implemented.

Considerable additional work is required to continue the effort begun in tying together the joint effects of correlation and track update in the tracking function. Many of the problems discussed in [2] still remain unsolved.

VI. REFERENCES

1. N.E. Nahi, "Optimal Recursive Estimation with Uncertain Observation," IEEE Trans. Information Theory, Vol, IT-15, pp. 457-462, July 1969.
2. R.A. Singer and J.J. Stein, "An Optimal Tracking Filter for Processing Sensor Data of Imprecisely Determined Origin in Surveillance Systems," IEEE Conference on Decision and Control, Miami Beach, Florida, December, 1971.
3. R.E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, J. Basic Engrg, Vol. 82, pp 34-45, March 1960
4. T.R. Benedict and G.W. Bordner, "Synthesis of an Optimal Set of Radar Track-While-Scan Smoothing Equations," IRE Trans Automatic Control, Vol. AC-7, pp 27-32, July 1962
5. R.A. Singer and K.W. Behnke, "Real-Time Tracking Filter Evaluation and Selection for Tactical Applications," IEEE Trans Aerospace and Electronic Systems, Vol AES-7, pp 100-111, January 1971
6. R.P. Wishner, R.E. Larson, and M. Athans, "Status of Radar Tracking Algorithms," Proceedings of the Symposium on Nonlinear Estimation Theory and its Applications, San Diego, pp 32-54, September, 1970
7. R.K. Mehra, "On the Identification of Variances and Adaptive Kalman Filtering," IEEE Trans Automatic Control, Vol AC-15, pp 175-184, April 1970
8. R.G. Sea, "An Efficient Suboptimal Decision Procedure for Associating Sensor Data with Stored Tracks in Surveillance Systems," IEEE Conference on Decision and Control, Miami Beach, Florida, December 1971

APPENDIX: DERIVATION OF CORRELATION DEPENDENT TERMS $A(k,G)$, $R_2(k,G)$, and $P_2(k,G)$:

Letting cor denote "correct correlation" and det denote "detection," and after dropping indices, it follows from (31) and (33) that

$$A \triangleq \int\int v\, e'\, p\,(v,\, e\,|\,\text{cor, det})\, dv\, de \tag{A.1}$$

$$R_2 \triangleq \int\int vv'\, p\,(v,\, e\,|\,\text{cor, det})\, dv\, de \tag{A.2}$$

$$P_2 \triangleq \int\int e\, e'\, p\,(v,\, e\,|\,\text{cor, det})\, dv\, de \tag{A.3}$$

The limits of integration are infinity, although $p\,(v,\, e\,|\,\text{cor, det}) = 0$ when $He + v$ is outside the gate $\tilde{G}$. Applying Bayes rule, and noting that v and e are independent when conditioned only on detection, it is found that

$$A = \frac{1}{P\,(\text{cor/det})} \int p\,(e) \left[\int P\,(\text{cor/det},\, v,\, e)\, p\,(v)\, v\, dv \right] e'\, de \tag{A.4}$$

$$R_2 = \frac{1}{P\,(\text{cor/det})} \int p\,(e) \left[\int P\,(\text{cor/det},\, v,\, e)\, p\,(v)\, vv'\, dv \right] de \tag{A.5}$$

$$P_2 = \frac{1}{P\,(\mathrm{cor/det})} \int p\,(e) \left[\int P\,(\mathrm{cor/det},\, v,\, e)\; p\,(v)\, dv \right] e\, e'\, de \tag{A.6}$$

These equations can be solved explicitly since the quantities v and e are zero mean, independent, Gaussian random variables with covariance matrices R(k) and P(k/k-1), respectively, and since, from (17),

$$P\,(\mathrm{cor/det},\, v,\, e) = \begin{cases} \exp\left[\dfrac{-\hat{n}_{TF}\, C_N\, \pi\, \sqrt{|\hat{V}|}\, \left([He + v]'\, \hat{V}^{-1}\, [He + v] \right)^{N/2}}{2\hat{\sigma}_1 \ldots 2\hat{\sigma}_N} \right] & \text{if } He + v \in \tilde{G} \\ 0 \text{ if } He + v \notin \tilde{G} \end{cases} \tag{A.7}$$

$R_W(k)$: In order for the residual corresponding to the closest incorrect return to be declared correlated, it is necessary that there be no other returns with residuals in the region $\tilde{G} \cap D\,(w)$ where $D(w) = \{ \tilde{y} \mid \tilde{y}'\, \hat{V}^{-1}\, \tilde{y} \le w'\, \hat{V}^{-1}\, w \}$. Using Bayes' rule

p(w | w is residual of an incorrectly correlated return) =

$$p\,(w \mid \text{no other returns in } \tilde{G} \cap D(w)) = \frac{P\left(\begin{matrix} \text{no other returns} \\ \text{in } \tilde{G} \cap D\,(w) \end{matrix} \,\middle|\, w \right) p\,(w)}{P\,(\text{no other returns in } \tilde{G} \cap D\,(w))} \tag{A.8}$$

But

$$P\left(\begin{matrix} \text{no other returns} \\ \text{in } \tilde{G} \cap D(w) \end{matrix} \,\middle|\, w \right) = \left[1 - \frac{P_D}{(2\pi)^{N/2}\, \sqrt{|V|}} \int_{\tilde{G} \cap D(w)} e^{-\frac{1}{2} \tilde{y}'\, V^{-1}\, \tilde{y}}\, d\tilde{y} \right]$$

$$\exp\left[\frac{-\hat{n}_{TF}}{2\hat{\sigma}_1 \ldots 2\hat{\sigma}_N} \int_{\tilde{G} \cap D(w)} d\tilde{y} \right] \tag{A.9}$$

and

$$p\,(w) = \frac{1}{2K_1 \hat{\sigma}_1 \ldots 2K_N \hat{\sigma}_N} \text{ for } w \in \tilde{G}, = 0 \text{ for } w \notin \tilde{G} \tag{A.10}$$

From these equations, the appropriate conditional p.d.f. on w is obtained; hence, the desired covariance $R_W(k)$ may be evaluated. A useful approximation replaces $\tilde{G} \cap D(w)$ by D(w) since D(w) is contained in $\tilde{G}$ except where w is near the corners of $\tilde{G}$.

P(k/k-1, no update), P(k/k-1, update): Using

$$P(k/k-1,\ \text{no update}) \triangleq E\left[e\,(k \mid k-1)\; e'\,(k/k-1) \mid \text{no update} \right], \tag{A.11}$$

it follows that by finding the conditional p.d.f. on e, p (e | no update), using Bayes' rule, by using an expression for P(no update | e) derivable from [8], by recalling that e is a zero mean gaussian random variable with covariance P(k/k-1), and by using P (no update) = $1 - P_G$, then P(k/k-1, no update) may be found directly. The expression for P(k/k-1, update) is then simply

$$P(k/k-1,\ \text{update}) = \left[P(k/k-1) - P(k/k-1,\ \text{no update})\,(1 - P_G) \right] / P_G \tag{A.12}$$

B(k, G): To show that for the correlation algorithm given here, $B(k,G) = E\left[w(k)\, e'(k/k-1) \mid \text{incorrect update}\right]$ is zero, Bayes' rule is used to obtain the conditional joint p.d.f. on w and e:

$$p\,(w, e \mid \text{incorrect update}) = \frac{P\,(\text{incorrect update} \mid e,\ w)\ p(e)p(w)}{P\,(\text{incorrect update})} \tag{A.13}$$

Using the uniform distribution of incorrect plots and the symmetry of the gate ($w = 0$ corresponds to center of gate), the p.d.f. $p(w)$ is seen to be an even function of w. Further, the probability of incorrect update, given e and w, is an even function of w because the statistical distance test involves an even function of w, namely $w'\,\hat{V}^{-1} w$. Thus $p\,(w,\ e \mid \text{incorrect update})$ is an even function of w, and so the expected value of we', given incorrect update, must be zero.

Special case: N = 2, infinite gate

For the two dimensional measurement case, with infinite gate, used in the example, the general expressions for A, R_2, P_2, and R_w may be evaluated, after considerable manipulation, to be:

$$A = -R\,\hat{V}_s^{-1}\,H\,P_2,\quad R_2 = R - R\hat{V}_s^{-1}\left[\hat{V}_s - H\,P_2 H'\right]\hat{V}_s^{-1}\,R$$

$$P_2 = \left[P\,(k/k-1)^{-1} + H'\,\hat{V}_s^{-1}\,H\right]^{-1},\quad R_w = h\,(V,\ \hat{V},\ P_D,\ n_{TF},\ \hat{n}_{TF})\ \hat{V}$$

where $\hat{V}_o = \hat{V}\left[\frac{\pi}{2}\,\hat{n}_{TF}\,\sqrt{1 - \hat{\rho}^2}\right]^{-1}$, $\hat{V}_s = \hat{V}_o + R$

and where h is a known, complicated, function of the variables indicated.

SOME NEW RESULTS IN NON-GAUSSIAN DISCRETE-TIME ESTIMATION THEORY RECURSIVE POLYNOMIAL ESTIMATORS

B. D. JENSEN
Bell Telephone Laboratories, Incorporated
Whippany, New Jersey 07981

ABSTRACT

This paper demonstrates how recursive polynomial estimators, polynomial functions of the observation data, can be derived for a class of non-Gaussian discrete-time estimation problems. This class includes estimating the state vector of a linear non-Gaussian dynamic system from linear and nonlinear observations, estimating the moments of a random vector from independent samples, joint signal detection estimation, and system noise moment estimation for adaptive filtering. The theory is illustrated with numerous scalar examples.

INTRODUCTION

It is well known that the minimum mean-square error (MMSE) estimate of a discrete parameter random vector x_n, as an arbitrary function of a statistically related sequence of random vectors, $\Omega_m = \{y_i; i=1,2,\dots,m\}$, is the conditional expectation of x_n given Ω_m. If x_n and Ω_m are jointly Gaussian this estimate is the orthogonal projection of x_n onto the linear subspace spanned by the observations. On the other hand, if x_n and Ω_m are not jointly Gaussian, the conditional mean is, in general, a highly nonlinear function and can only be computed efficiently for a relatively few problems and approximations are required. In this paper we consider polynomial approximations of the conditional mean, and we demonstrate how recursive estimation algorithms can be derived for a wide variety of non-Gaussian estimation problems. Although these estimators are not the unrestricted MMSE estimators, they are the exact MMSE estimators of a specified structure, and in many cases they are significantly better than existing approximations.

The polynomial estimation theory is developed in three steps. First we present a set of equations, Theorem 1 in the next section, that specify the orthogonal projection of a zero-mean, second-order, random vector, ω_n, onto the linear subspace spanned by the sequence of zero-mean, second-order, random vectors, $\{z_i; i=1,2,\dots,\}$, i.e., the MMSE nonrecursive linear estimator. For the problems we consider here, the existing polynomial estimation theory [1,2,3,4] is encompassed by this theorem if the appropriate polynomial terms are included in the definition of the vectors ω_j and z_j. However, our interest here is in recursive estimators, and we use Theorem 1 as a tool in the second step of our development to derive the MMSE linear recursive estimator (only the filtering estimator is given) for a certain linear system. This result is given in Theorem 2 of the third section. To our knowledge all polynomial estimation problems that can be solved in recursive form can be formulated as special cases of this linear system and we refer to this formulation as the Linear System Representation or simply the LSR. Finally the third and possibly the most significant step is the construction of the LSR for particular polynomial estimation problems. We illustrate this step in the fourth section by considering a wide variety of scalar examples. Some numerical results that demonstrate the usefulness of polynomial filters are also given there. These examples do not exhaust the

the theory but merely indicate its broad scope. Furthermore, the techniques are easily extended to the vector case by applying the manipulations to the individual components.

ORTHOGONAL PROJECTION

Theorem 1 completely solves the linear estimation problem. Essentially it consists of a concise statement of some of the orthogonal projection ideas introduced by Kolmogorov and advanced by Doob and by Wiener and Masani [5]. The equations in this form were apparently derived and used by the author [7] and Aasnaes [6] at about the same time. Numerous other papers contain equivalent equations for less general estimation problems or equivalent equations in an inconvenient form.

Theorem 1 The orthogonal projection of the zero-mean, second-order, random vector ω_n onto the linear subspace spanned by the zero-mean, second-order, sequence of random vectors $\{z_i;\ i = 1,2,\ldots,m\}$ is $\hat{\omega}_{n/m}$ where

$$\hat{\omega}_{n/m} = \sum_{i=1}^{m} P_{\omega\tilde{z}}(n,i/i-1)P^{+}_{z\tilde{z}}(i,i/i-1)\tilde{z}_i \qquad (1)^{*}$$

$$\tilde{z}_n = z_n - \sum_{i=1}^{n-1} P_{z\tilde{z}}(n,i/i-1)P^{+}_{z\tilde{z}}(i,i/i-1)\tilde{z}_i \qquad (2)$$

and

$$P_{\alpha\tilde{\beta}}(n,m/k) = E\left(\alpha_n\beta_m'\right) - \sum_{i=1}^{k} P_{\alpha\tilde{z}}(n,i/i-1)P^{+}_{z\tilde{z}}(i,i/i-1)P'_{\beta\tilde{z}}(m,i/i-1); \qquad (3)$$

furthermore, the mean-square estimation error is given by $P_{x\tilde{x}}(n,n/n)$.

The proof of this theorem is straightforward [7]. Note that $\tilde{z}_n \equiv z_n - \hat{z}_{n/n-1}$ is the innovations sequence and

$$P_{\alpha\tilde{\beta}}(n,m/k) \equiv E\left[\alpha_n(\beta_m - \hat{\beta}_{m/k})'\right].$$

The following Corollary of Theorem 1 is extremely useful for deriving recursive estimators.

Corollary 1 - Linearity Property

If

$$E\left(\alpha_n\beta_m'\right) = A_r E\left(a_r\beta_m'\right) + B_s E\left(b_s\beta_m'\right)$$

then

---

[*] $P^{+}_{z\tilde{z}}(\ )$ is a generalized inverse of $P_{z\tilde{z}}(\ )$. For all of the problems we consider the inverse of $P_{z\tilde{z}}(\ )$ exists so that $P^{+}_{z\tilde{z}}(\ ) = P^{-1}_{z\tilde{z}}(\ )$. Also note that summations of the form $\sum_{i=\ell}^{k}(\ )i$ are zero for $k < \ell$.

$$P_{\alpha\tilde{\beta}}(n,m/k) = A_r P_{a\tilde{\beta}}(r,m/k) + B_s P_{b\tilde{\beta}}(s,m/k)$$

which implies from (1) with $\beta_n = z_n$ that

$$\hat{\alpha}_{n/k} = A_r \hat{a}_{r/k} + B_s \hat{b}_{s/k}.$$

Proof: See [7]

LINEAR SYSTEM REPRESENTATION

The general LSR is given by equations (4) below, and the MMSE linear filter $(m = n)$ for this system is given by Theorem 2. Although we only consider the filtering problem here, Theorem 1 is valid for an arbitrary relationship between m and n and, as indicated in [7], it can be used to derive other types of estimators as well.

To construct an LSR for a particular problem it is necessary to select a polynomial observation vector (POV), Y_i, such that the components of Y_i are the desired polynomial functions of the components of the actual observations $\{y_k;\ k = i, i-1, \ldots, i-\lambda\}$. Then for the problems of interest here the vector to be estimated, x_n, is related to the POV by the following linear system.

General LSR

$$\omega_{n+1} = \psi(n+1,n)\omega_n + G_n(\omega_n)u_n \tag{4a}$$

$$z_n = H_n\omega_n + F_n(\omega_n)v_n \tag{4b}$$

$$z_n = Y_n - \overline{Y}_n, \qquad x_n = L\omega_n + \overline{x}_n \tag{4c}$$

where $E(\ ) \equiv (\overline{\ \ })$ and the following three conditions are satisfied:

1. ω_o, $\{u_i;\ i = 0,1,\ldots\}$, and $\{v_i;\ i = 1,2,\ldots\}$ are zero-mean random vector with covariances

   $E\left(\omega_o u_i'\right) = 0, \quad E\left(\omega_o v_i'\right) = 0, \quad E\left(u_i v_j'\right) = 0, \quad E\left(\omega_o \omega_o'\right) = P_o,$

   $E\left(u_i u_j'\right) = Q_i \delta_{ij}, \quad$ and $\quad E\left(\mathbf{v}_i v_j'\right) = R_{ij}, \quad$ where $\quad R_{ij} = 0 \quad$ for $\quad |i-j| > \lambda.$

2. $G_n(\omega_n)$ and $F_n(\omega_n)$ are random functions of the system state that are independent of u_i for $i \geq n$ and v_i for all i, and the quantities $E\left[G_n(\omega_n)Q_n G_n'(\omega_n)\right] = Q_n^*$ and $E\left[F_i(\omega_i)R_{ij}F_j'(\omega_j)\right] = R_{ij}^*$ can always be computed.

3. $\psi(n+1,n)$ and H_n are deterministic matrices, L is a constant matrix, and λ is a nonnegative integer.

Theorem 2 The MMSE linear filter for the general LSR is

$$\hat{\omega}_{o/o} = 0, \quad \hat{v}^*_{1/o} = 0, \quad \alpha \equiv \max(n-\lambda,0) \tag{5a}$$

$$\hat{x}_{n/n} = L\hat{\omega}_{n/n} + \bar{x}_n \tag{5b}$$

$$\hat{\omega}_{n/n-1} = \psi(n,n-1)\hat{\omega}_{n-1/n-1} \tag{5c}$$

$$\tilde{z}_n = z_n - H_n\hat{\omega}_{n/n-1} - \hat{v}^*_{n/n-1} \tag{5d}$$

$$\hat{\omega}_{n/n} = \hat{\omega}_{n/n-1} + K_n\tilde{z}_n \tag{5e}$$

$$\hat{v}^*_{n/n-1} = \sum_{i=\alpha}^{n-1} S(n,i)N_i^{-1}\,\tilde{z}_i \tag{5f}$$

where the filter gains can be computed off-line or in real time from

$$P_{n/n-1} = \psi(n,n-1)P_{n-1/n-1}\psi'(n,n-1) + Q^*_{n-1}, \quad P_{o/o} = P_o \tag{5g}$$

$$S(n,i) = R^*_{n,i} - \sum_{j=\alpha}^{i-1} S(n,j)\left[H_i\psi(i,j)K_j + S(i,j)N_j^{-1}\right]' \tag{5h}$$

$$N_n = H_nP_{n/n-1}H_n' + S(n,n) - \sum_{i=\alpha}^{n-1} H_n\psi(n,i)K_iS'(n,i) \tag{5i}$$

$$K_n = \left[P_{n/n-1}H_n' - \sum_{i=\alpha}^{n-1} \psi(n,i)K_iS'(n,i)\right]N_n^{-1} \tag{5j}$$

$$P_{n/n} = P_{n/n-1} - K_nN_nK_n' \tag{5k}$$

and the mean-square estimation error is

$$E\left[(x_n-\hat{x}_{n/n})(x_n-\hat{x}_{n/n})'\right] = LP_{n/n}L'. \tag{5l}$$

Proof: Equations (5b), (5c), and (5d) follow directly from Corollary 1 since

$$E\left(\omega_n z_i'\right) = \psi(n,k)E\left(\omega_k z_i'\right), \qquad i \le k \le n \tag{6}$$

and

$$E\left(z_n z_i'\right) = H_nE\left(\omega_n z_i'\right) + E\left(v_n^* z_i'\right) \tag{7}$$

where

$$v_n^* = F_n(\omega_n)v_n.$$

Equations (5e) and (5f) follow from (1) by defining

$$N_n \equiv P_{z\tilde{z}}(n,n/n-1), \quad K_n \equiv P_{\omega\tilde{z}}(n,n/n-1)N_n^{-1}, \quad S(n,i) \equiv P_{v^*\tilde{z}}(n,i/i-1)$$

and noting from Corollary 1 that since

$$E\left(v_n^* z_i'\right) = 0 \text{ for } |n-i| > \lambda, \quad S(n,i) = 0 \text{ for } |n-i| > \lambda.$$

Equation (5g) follows from (3) and the definition $P_{\omega\tilde{\omega}}(n,n/k) \equiv P_{n/k}$ by setting

$$E\left(\omega_n\omega_n'\right) = \psi(n,n-1)E\left(\omega_{n-1}\omega_{n-1}'\right)\psi'(n,n-1) + Q_{n-1}^*$$

and noting from (6) and Corollary 1 that

$$P_{\omega\tilde{z}}(n,i/i-1) = \psi(n,n-1)P_{\omega\tilde{z}}(n-1,i/i-1) \quad \text{for} \quad i \leq n-1. \tag{8}$$

From (8) and the definition of K_i we can write

$$P_{\omega\tilde{z}}(n,i/i-1)N_i^{-1} = \psi(n,i)K_i \tag{9}$$

and from (7) and Corollary 1 we have

$$P_{z\tilde{z}}(n,i/i-1) = H_nP_{\omega\tilde{z}}(n,i/i-1) + S(n,i). \tag{10}$$

Now since $E\left(\omega_nz_n'\right) = E(\omega_n\omega_n')H_n'$ we can use (9), (10) and (3) to write

$$P_{\omega\tilde{z}}(n,n/n-1) = P_{n/n-1}H_n' - \sum_{i=\alpha}^{n-1} \psi(n,i)K_iS'(n,i) \tag{11}$$

Equation (5h) follows from (3) and the definitions; equation (5i) results from (10) and (11), and equation (5j) follows directly from (11) and the definition of K_i. Equation (5k) is a restatement of (3) with $\alpha = \beta = \omega$ and $n = m = k$, and equation (5l) is obvious from (5b). QED

Note that if Y_i is only a function of y_i then $\lambda = 0$ and all of the summations in (5) are zero by definition.

APPLICATIONS

A. Linear Dynamic System with Non-Gaussian Statistics

Here we consider the problem of estimating the state, x_n, of the linear scalar system

$$x_{n+1} = \phi_nx_n + \zeta_n; \qquad E\left(x_o^k\right) = \overline{x_o^k}, \quad E\left(\zeta_i^k\right) = q_i^{(k)} \tag{12a}$$

from the noisy measurements

$$y_n = x_n + \eta_n; \qquad E\left(\eta_i^k\right) = r_i^{(k)} \tag{12b}$$

where the random variables $\{x_o;\ \zeta_i,\ i=0,\ldots;\eta_i,\ i=1,2,\ldots\}$ are zero-mean and independent. The POV must be determined separately for each specific problem by weighing the MSE reduction against the increase in computational complexity. To illustrate the required manipulations we will derive the LSR for (12) with the POV $Y_n = \left(y_n,y_n^2,y_n^3,y_ny_{n-1},y_n^2y_{n-1}\right)'$. Directly from (12b) we can write

$$z_n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 3r_n^{(2)} & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & r_n^{(2)} \end{bmatrix} \omega_n + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2x_n & 1 & 0 & 0 & 0 & 0 \\ 3x_n^2 & 3x_n & 1 & 0 & 0 & 0 \\ x_{n-1} & 0 & 0 & x_n & 1 & 0 \\ 2x_n x_{n-1} & x_{n-1} & 0 & x_n^2 & 2x_n & 1 \end{bmatrix} v_n \tag{13a}$$

where $z_n \equiv Y_n - \overline{Y}_n$,

$$\omega_n \equiv \left[x_n, \left(x_n^2 - \overline{x_n^2}\right), \left(x_n^3 - \overline{x_n^3}\right), \left(x_n x_{n-1} - \overline{x_n x_{n-1}}\right), \left(x_n^2 x_{n-1} - \overline{x_n^2 x_{n-1}}\right), x_{n-1}\right]',$$

$$v_n \equiv \left[\eta_n, \left(\eta_n^2 - r_n^{(2)}\right), \left(\eta_n^3 - r_n^{(3)}\right), \eta_{n-1}, \eta_n \eta_{n-1}, \eta_n^2 \eta_{n-1}\right]',$$

and we have added and subtracted the vector $\left[0, 0, 3x_n r_n^{(2)}, 0, x_{n-1} r_n^{(2)}\right]'$ to get the desired form. Similarly from the definition of ω_n and equation (12a) we can write

$$\omega_{n+1} = \begin{bmatrix} \phi_n & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_n^2 & 0 & 0 & 0 & 0 \\ 3\phi_n q_n^{(2)} & 0 & \phi_n^3 & 0 & 0 & 0 \\ 0 & \phi_n & 0 & 0 & 0 & 0 \\ q_n^{(2)} & 0 & \phi_n^2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \omega_n + \begin{bmatrix} 1 & 0 & 0 \\ 2\phi_n x_n & 1 & 0 \\ 3\phi_n^2 x_n^2 & 3\phi_n x_n & 1 \\ x_n & 0 & 0 \\ 2\phi_n x_n^2 & x_n & 0 \\ 0 & 0 & 0 \end{bmatrix} u_n \tag{13b}$$

where $u_n = \left[\zeta_n, \left(\zeta_n^2 - q_n^{(2)}\right), \left(\zeta_n^3 - q_n^{(3)}\right)\right]'$.

The system (13) is the LSR for the problem (12) with $\lambda = 1$ and $L = (1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0)$. This particular POV requires two observations, but if desired it can be started from the first observation by simply defining H_1 and $F_1(\omega_1)$ as the first three rows in (13a). The quantities Q_n^*, R_n^*, and P_o are easily computed from their definitions and the time-varying moments of x_n which can be computed from (12a).

Clearly, the filter algorithm given by Theorem 2 is seldom in the most efficient form from a computational standpoint. Considerable savings usually result by partitioning and other manipulations that take advantage of the large number of zeros in the various filter matrices. For a more complete discussion of the various aspects of polynomial filters for linear systems we refer the reader to [7].

B. Nonlinear Observations

With the model (12a) for x_n, consider the case where the observations are of the form

$$y_n = p(x_n) + \eta_n ; \qquad E\left(\eta_i^k\right) = r_i^{(k)} \tag{14}$$

where $p(x_n)$ is a polynomial in x_n and the random variables are statistically related as in problem A. Clearly this problem is a straightforward generalization of problem A and can be solved in the same manner by simply including in ω_n all of the polynomial terms of x_n that are generated by the desired POV. However, as opposed to problems of the type A, the MMSE linear estimate is not the conditional mean when the noise sequences and the initial state are Gaussian. Furthermore, the MMSE linear filter is significantly different from what is often referred to as the "optimal linear filter", i.e., a Kalman filter derived by a first order approximation. Again we refer the reader to [7] for more details and numerical results.

C. Joint Detection Estimation

A problem that has received considerable attention in the last few years is that of joint detection estimation or estimation with uncertain observations. One of the problems that fall in this class is the following. Again we let (12a) be the signal model, but this time the observations are of the form

$$y_n = \gamma x_n + \eta_n , \qquad E\left(\eta_i^k\right) = r_i^{(k)} \tag{15}$$

where γ is a 0-1 random variable such that $\mathrm{Prob}(\gamma = 1 \ \text{ for } \ n = 1,2,\ldots) = \beta$ and $\mathrm{Prob}(\gamma = 0 \ \text{ for } \ n = 1,2,\ldots) = 1 - \beta$. For this problem it is possible to derive the exact MMSE estimator for both x_n and γ, but this solution is quite complicated, particularly in the non-Gaussian case. Nahi [8]* derived a simple approximation for estimating x_n, namely, the MMSE linear filter. The MMSE quadratic filter, $Y_n = (y_n, y_n^2)'$, can be derived by applying Theorem 2 to the following LSR.

$$\omega_{n+1} = \begin{pmatrix} \phi_n & 0 & 0 & 0 \\ 0 & \phi_n & 0 & 0 \\ 0 & 0 & \phi_n^2 & q_n^{(2)} \\ 0 & 0 & 0 & 1 \end{pmatrix} \omega_n + \begin{pmatrix} 1 & 0 \\ \gamma & 0 \\ 2\gamma^2 \phi_n x_n & \gamma^2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \zeta_n \\ \\ \zeta_n^2 - q_n^{(2)} \end{pmatrix} \tag{16a}$$

$$\omega_n \equiv \left[(x_n - \bar{x}_n), (\gamma x_n - \beta \bar{x}_n), \left(\gamma^2 x_n^2 - \beta \overline{x_n^2}\right), (\gamma^2 - \beta) \right]'$$

$$z_n = Y_n - \bar{Y}_n = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \omega_n + \begin{pmatrix} 1 & 0 \\ 2\gamma_n x_n & 1 \end{pmatrix} \begin{pmatrix} \eta_n \\ \eta_n^2 - r_n^{(2)} \end{pmatrix} \tag{16b}$$

$$\lambda = 0 \qquad \text{and} \qquad L = (1 \quad 0 \quad 0 \quad 0).$$

These equations follow from the definitions by adding and subtracting $\gamma^2 q_n^{(2)}$ to the third state equation. Note that we have assumed $\bar{x}_o \neq 0$ so that x_n

---

[*] Nahi's paper contains an error. The correct solution is given in [7].

would be correlated with y_n^2 even for Gaussian x_n. Otherwise for Gaussian x_n a higher order POV would be required.

Table 1 below illustrates the significant improvement in estimation accuracy that can be achieved by the quadratic filter. Also shown is the MSE for a Kalman filter that is "told" the value of γ.

Table 1 Estimation with Uncertain Observations System:

$\beta = .95, \quad \phi_n = 1.0, \quad q^{(i)} = 0, \quad \eta_i$ and x_o Gaussian

with $r_n^{(2)} = 1.0, \quad \overline{x}_o = 10$, and $\sigma_x^2 = 1.0$

	Time (n)	1	2	3	4	5	10	50	100
MSE	Linear	.87	.85	.85	.85	.85	.84	.84	.84
	Quadratic	.55	.40	.32	.27	.26	.16	.089	.079
	Kalman given γ	.53	.37	.29	.24	.21	.14	.069	.059

D. Moment Estimation

In [7] we considered the problem of estimating the k^{th} time-varying moment of a scalar random variable from a sequence of conditionally independent observations, $\{y_i;\ i = 1,2,\dots,n\}$, by using a POV of the form $Y_n = y_n^k$. This theory can be extended in a number of ways. We have, for example, derived recursive algorithms for simultaneously estimating several moments, for estimating the moments of correlated sequences, and for estimating a given moment with a higher order polynomial estimator. This latter extension is illustrated here.

The problem considered is that of estimating the constant variance, m, of a Gaussian random variable, y_i. The MMSE quadratic estimator is given in [7]. Here we will present the LSR for the MMSE quartic estimator and present some numerical results comparing the two polynomial filters with the maximum likelihood estimator.

For y_i zero mean, m is defined by $m = E\left(y_i^2/m\right)$; and for y_i Gaussian, we have

$$E\left(y_i^{2k}\right) = E\left[E\left(y_i^{2k}/m\right)\right] = (2k-1)!\left(\overline{m^k}\right).$$

Consequently, by adding and subtracting $(2k-1)!\,m^k$ we can write

$$y_i^{2k} - \overline{y_i^{2k}} = (2k-1)!\left(m^k - \overline{m^k}\right) + \left[y_i^{2k} - (2k-1)!\,m^k\right].$$

Therefore, by defining $\omega_n \equiv \left[(m-\overline{m}),\ \left(3m^2 - 3\overline{m^2}\right)\right]'$ and $v_n \equiv \left[\left(y_n^2 - m\right),\left(y_n^4 - 3m^2\right)\right]'$ we have the following LSR for a POV of the form $Y_n = \left(y_n^2, y_n^4\right)'$

$$\omega_{n+1} = \omega_n \tag{17a}$$

$$z_n = \omega_n + v_n \tag{17b}$$

Unusual as it may seem, equations (17) satisfy the requirements of a LSR since, in particular, $E\left(\omega_i v_j'\right) = 0$. Table 2 compares the square-root of the MSE for the quadratic and quartic filters with the standard deviation of the maximum likelihood estimate (maximum likelihood for nonrandom m) for a particular case.

Table 2 Moment Estimation

m: Gamma with $\bar{m} = 10$ and $E(m-\bar{m})^2 = 1.0$

n	0	5	10	50	100	200
Quadratic	1.0	.99	.98	.90	.82	.71
Quartic	1.0	.94	.88	.64	.51	.39
Maximum Likelihood	14.2	5.8	4.3	1.99	1.41	1.00

E. Linear System Noise Moment Estimation--Adaptive Filtering

The last problem we consider is that of estimating the variance of the plant and observation noise for the standard linear model (12). Thus we assume $E\left(\zeta_n^2/q_n\right) = q_n$ and $E\left(\eta_n^2/r_n\right) = r_n$ where

$$q_{n+1} = a_n q_n + \xi_n, \quad E(\xi_i \xi_j) = s_i \delta_{ij} \tag{18a}$$

$$r_{n+1} = b_n r_n + \mu_n, \quad E(\mu_i \mu_j) = t_i \delta_{ij} \tag{18b}$$

and x_o, ξ_i, μ_i, (ζ_i/q_i), and (η_i/r_i) are independent zero mean random variables. The MMSE quadratic filter can be derived from Theorem 2 and the following LSR.

$$\omega_{n+1} = \begin{pmatrix} \phi_n^2 & 1 & 0 \\ 0 & a_n & 0 \\ 0 & 0 & b_n \end{pmatrix} \omega_n + \begin{pmatrix} 2\phi_n x_n & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \zeta_n \\ \zeta_n^2 - q_n \\ \xi_n \\ \eta_n \end{pmatrix} \tag{19a}$$

$$z_n = y_n^2 - \overline{y_n^2} = (1 \quad 0 \quad 1)\,\omega_n + (2x_n \quad 1) \begin{pmatrix} \eta_n \\ \eta_n^2 - r_n \end{pmatrix} \tag{19b}$$

where $\omega_n = \left[\left(x_n^2 - \overline{x_n^2}\right), (q_n - \bar{q}_n), (r_n - \bar{r}_n)\right]'$, $\lambda = 1$ and for estimating q_n and r_n, $L = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Since the estimates of q_n and r_n can be computed on line, they can be used in an adaptive Kalman filter. Due to its generality, this approach to adaptive filtering has several advantages over the existing approaches.

CONCLUSION

We have shown how recursive, minimum mean-square error, polynomial estimators can be derived for a broad class of non-Gaussian estimation problems. As we have demonstrated, the practical applications of this theory are numerous.

REFERENCES

1. P. Masani and N. Wiener, "Non-Linear Prediction" Probability and Statistics: The Herald Cramer Volume, Wiley, 1959.

2. H. V. Balakrishan, "On a Class of Nonlinear Estimation Problems," IEEE Trans. on Information Theory, IT-10(4), 314-320, October 1964.

3. W. H. Kroy, Jr., "Identification of Linear Stochastic Dynamical Systems by Nonlinear Weighting of Noisy Meassurements," UCLA Report No. 66-57, October 1966.

4. A. R. Stubberud and W. K. Masenten, "Polynomial Estimators," 1970 IEEE Symposium on Adaptive Processes - Decision and Control, The University of Texas at Austin, December 7-9, 1970.

5. P. Masani and N. Wiener, "The Prediction Theory of Multivariable Stochastic Process, I," Acta. Math. 98, 1957.

6. H. B. Aasnaes, "Estimation, Prediction and Identification in Linear Discrete-Time Systems Using the Innovations Concept," Ph. D. Dissertation, Dept. of EE, Stanford University, 1970.

7. B. D. Jensen, "New Results and Techniques in Discrete-Time Recursive Estimation Theory," Ph. D. Dissertation, Systems Science, Polytechnic Institute of Brooklyn, 1971.

8. N. E. Nahi, "Optimal Recursive Estimation with Uncertain Observation," IEEE Trans. on Information Theory, IT-15(4), July 1969.

ESTIMATION FOR CYCLOSTATIONARY RANDOM PROCESSES

L. E. FRANKS and W. A. GARDNER
Department of Electrical Engineering
University of Massachusetts
Amherst, Massachusetts

ABSTRACT

The extent to which the time origin or "phase" of a cyclostationary process is known has a direct bearing on the design of the linear, periodically-varying filter used for detection or estimation. In this paper, we examine the continuous waveform estimation problem and compare performance with that of the best time-invariant filter. Illustrative examples involving various synchronous signal multiplexing schemes are presented.

I. PROPERTIES OF CYCLOSTATIONARY PROCESSES

A random signal process which is produced by a periodic sampling, scanning, or synchronous multiplexing operation can often be modeled by a process which exhibits periodicity in its mean and autocorrelation functions. A process, $x(t)$, is said to be cyclostationary in the wide sense (with period T) if

$$\begin{aligned} E[x(t)] &= E[x(t+T)] \\ k_{xx}(s,t) &= E[x(s)x^*(t)] = k_{xx}(s+T,\ t+T) \end{aligned} \tag{1}$$

for all s and t. Bennett [1] introduced the term, "cyclostationary", to denote this class of processes in his treatment of synchronously timed pulse sequences used in digital data transmission. Other investigators[2]-[5] have used terms such as "periodically stationary", "periodically correlated", and "periodic nonstationary" to denote this same class.

We shall assume that the process under consideration can be represented in the form[5]

$$x(t) = \sum_{n=-\infty}^{\infty} a_n(t) \exp\left[j2\frac{\pi n t}{T}\right] \tag{2}$$

where choosing

$$\begin{aligned} a_n(t) &= \int_{-\infty}^{\infty} v(t-\tau) \exp\left[-j\frac{2\pi n\tau}{T}\right] x(\tau)d\tau \\ v(t) &= \frac{1}{\pi t} \sin\frac{\pi t}{T} \end{aligned} \tag{3}$$

makes the mean-squared value of the difference of both sides of (2) vanish.

Generation of the sequence of bandlimited random processes, $\{a_n(t)\}$, is illustrated in Fig. 1. If $x(t)$ is cyclostationary, then we can show that $\{a_n(t)\}$ is jointly wide-sense stationary and we define:

$$r_{nm}(s-t) = E[a_n(s)a_m^*(t)] \tag{4}$$

with the result that

$$k_{xx}(s,t) = \sum_n \sum_m r_{nm}(s-t) \exp\left[j \frac{2\pi}{T} (ns-mt)\right] \tag{5}$$

There are certain properties of the infinite-dimensional matrix, $\underline{r}(\cdot)$, or its Fourier transform $\underline{R}(\cdot)$ which will be useful in the estimation problem.

1.) $\underline{R}(f)$ is bandlimited to the frequency interval, $|f| \leq \frac{1}{2T}$. This is because the $\{a_n(t)\}$ are similarly bandlimited.

2.) $x(t)$ is a wide-sense stationary process if, and only if, $\underline{r}(\cdot)$is a diagonal matrix. This can be seen by direct substitution into (5).

3.) Let $x(t)$ be the result of a time-invariant filtering operation on a cyclostationary $x'(t)$, with representation $\underline{R}'(\cdot)$, given by

$$x(t) = \int_{-\infty}^{\infty} h(t-\tau)\, x'(\tau) d\tau$$

then we have

$$R_{nm}(f) = V(f)\, H(f+\tfrac{n}{T})\, H^*(f+\tfrac{m}{T})\, R'_{nm}(f) \tag{6}$$

Using this result with $H(\cdot)$ corresponding to an ideal lowpass filter with cutoff at W, we see that the bandlimited process, $x(t)$, can be represented by a finite matrix of order M where $M \leq 2\,TW$.

4.) Let $x(t)$ be derived from a cyclostationary $x'(t)$ where an uncertainty in the time origin or "phase" of the process has been introduced, i.e., let $x(t) = x'(t+\delta)$ where δ is a random variable with probability density function, $p_\delta(\cdot)$. The modification in the representation is expressed by

$$R_{nm}(f) = P_\delta\left(\frac{m-n}{T}\right) R'_{nm}(f) \tag{7}$$

where $P_\delta(\cdot)$ is the Fourier transform of $p_\delta(\cdot)$ so we know that

$$P_\delta\left(\frac{n}{T}\right) \leq P_\delta(0) = 1 \tag{8}$$

In particular, if δ is uniformly distributed over the interval $[0,T)$, then $P\left(\frac{n}{T}\right) = 0$ for $n \neq 0$ and $\underline{R}(f)$ is diagonal. Thus if a cyclostationary process is completely "phase randomized" it becomes a stationary process. Its autocorrelation function would be

$$\begin{aligned} k_{xx}(\tau) &= \sum_n r'_{nn}(\tau) \exp\left[j\, 2\pi \frac{n\tau}{T}\right] \\ &= \frac{1}{T}\int_0^T k_{x'x'}(t+\tau,t)\,dt \end{aligned} \tag{9}$$

which is simply a time-averaged version of the nonstationary correlation.

II. CONTINUOUS WAVEFORM ESTIMATION

The estimation problem we consider here is evaluation of the performance of the noncausal, periodically-variable filter which gives least mean-squared error reconstruction of a cyclostationary signal process with additive white noise having spectral density, N_o. The impulse response function, $h(\cdot,\cdot)$, for the filter which minimizes

$$I(t) = E\left[\left\{x(t) - \int_{-\infty}^{\infty} h(t,s)z(s)ds\right\}^2\right] \tag{10}$$

where $z(t)$ is signal plus white noise, must satisfy the orthogonality condition[6]

$$\int_{-\infty}^{\infty} k_{xx}(s,\sigma)h(t,\sigma)d\sigma + N_o h(t,s) = k_{xx}(s,t) \text{ for all } s \text{ and } t \tag{11}$$

Substitution of (11) into (10) and rearranging terms yields a compact expression for the minimum mean-squared error.

$$I_{min}(t) = N_o h(t,t) \tag{12}$$

Frequency-domain Method.

Using the representation scheme of (2), the optimum filter can be interpreted as an estimator of the jointly wide-sense stationary sequence, $\{a_n(t)\}$. This is accomplished by forming the corresponding sequence of processes for the received signal by means of the structure shown in Fig. 1.

The estimate of the n^{th} process is a linear combination of time-invariant operations on the received signal processes which is then multiplied by $\exp [j\ 2\pi \frac{nt}{T}]$ to form one component of the filter output. The structure of this estimator is shown in Fig. 2 and the overall filter is obtained as a cascade of the structures in Figs. 1 and 2. The filter is realized as a bank of input and output modulators interconnected by a matrix of time-invariant, bandlimited $\frac{1}{2T}$, filter paths. We can then write the impulse response as

$$h(t,s) = \sum_n \sum_k g_{nk}(t-s) \exp [j \frac{2\pi}{T} (nt-ks)] \tag{13}$$

Now using (13) and (5) in (11) and performing the integration, the coefficients of the exponentials in t and s are equated and the orthogonality condition is satisfied by a $\underline{g}(\cdot)$ matrix whose Fourier transform is the solution of

$$\underline{G}(f)\ [\underline{R}^a(f) + N_o\ \underline{I}] = \underline{R}^a(f);\ |f| \leq \frac{1}{2T} \tag{14}$$

where $\underline{R}^a(\cdot)$ denotes the conjugate transpose of $\underline{R}(\cdot)$.

As a measure of receiver performance, we use the value of $I_{min}(t)$ averaged over one period.

$$J_o = \frac{1}{T} \int_o^T I_{min}(t)dt = \frac{N_o}{T} \int_o^T h(t,t)dt \tag{15}$$

Using (13) in (15), the expression for optimum performance becomes

$$J_o = N_o \sum_n g_{nn}(o) = N_o \int_{-\infty}^{\infty} \mathrm{tr}\ \underline{G}(f)df \tag{16}$$

Optimum Time-invariant Filter.

It is of practical interest to evaluate the performance of the best time-invariant filter for a cyclostationary signal relative to that of the best periodically-variable filter. To do this we use h(t-s) in (10) and find the condition which minimizes the time-averaged value of I(t). The constraint of time-invariance is equivalently imposed by requiring that the $\underline{g}(\cdot)$ matrix for (13) be diagonal. The new orthogonality condition has the simple solution,

$$G_{nn}(f) = \frac{R_{nn}(f)}{R_{nn}(f) + N_o} \tag{17}$$

and since $\underline{G}(\cdot)$ is diagonal, the filter transfer function is

$$H(f) = \sum_n G_{nn}(f-\tfrac{n}{T}) \tag{18}$$

Taking into account the bandlimited nature of $R_{nn}(f)$ in (17), the sum in (18) can also be written as

$$H(f) = \frac{\sum_n R_{nn}(f-\frac{n}{T})}{\sum_n R_{nn}(f-\frac{n}{T}) + N_o} = \frac{K(f)}{K(f) + N_o} \tag{19}$$

The expression (19) is the familiar solution for the noncausal Wiener filter for a stationary signal with power spectral density K(f) in white noise[6]. Now referring to (9), we see that K(f) is actually the power spectral density for the (stationary) phase-randomized version of the cyclostationary process. Thus we make the interesting observation that the best time-invariant filter for the cyclostationary process is identical to the best filter for the phase-randomized version of the process. Furthermore, this filter is the result of simply disregarding the crosscorrelation among the $\{a_n(t)\}$ processes in making the estimation. The performance functional for the time-invariant filter is

$$J_1 = N_o h(o) = N_o \sum_n \int_{-\infty}^{\infty} \frac{R_{nn}(f-\frac{n}{T})}{R_{nn}(f-\frac{n}{T}) + N_o}\, df \tag{20}$$

Time-domain Method.

There is an alternative form of representation for the cyclostationary process which affords a simpler solution to the orthogonality condition in many cases of practical interest. This representation has the form

$$x(t) = \sum_i \sum_n \alpha_{ni}\, \phi_i\, (t-nT) \tag{21}$$

where the $\{\phi_i(t)\}$ are "doubly-orthogonal" in the sense that

$$\int_{-\infty}^{\infty} \phi_i(t-nT)\phi_j^*(t-mT)dt = \delta_{ij}\, \delta_{nm} \tag{22}$$

and the process is represented by the random variables

$$\alpha_{ni} = \int_{-\infty}^{\infty} x(t)\, \phi_i^*\, (t-nT)dt \tag{23}$$

This leads to a representation for the autocorrelation given by

$$k_{xx}(s,t) = \sum_{n}\sum_{m}\sum_{i}\sum_{j} A^{ij}_{n-m}\, \phi_i(s-nT)\, \phi^*_j(t-mT) \tag{24}$$

where

$$\begin{aligned} A^{ij}_{n-m} &= E\,[\alpha_{ni}\ \alpha^*_{mj}] \\ &= \int\!\!\int_{-\infty}^{\infty} k_{xx}(t+nT-mT,\tau)\ \phi^*_i(t)\phi_j(t)\,dt\,d\tau \end{aligned} \tag{25}$$

The optimum filter for the continuous waveform estimation problem has a similar representation.

$$h(t,s) = \sum_{\ell}\sum_{k}\sum_{p}\sum_{q} H^{qp}_{\ell-k}\ \phi_q(s-\ell T)\ \phi^*_p(t-kT) \tag{26}$$

Substituting (26) and (24) into the orthogonality condition (11) and performing the indicated integration by making use of (22), we obtain an equation whose terms have the same form as (24) and (26). Equating coefficients in this equation, the orthogonality condition is satisfied by the solution of

$$\sum_{m} \underline{A}_{n-m}\,\underline{H}_m + N_o\,\underline{H}_n = \underline{A}_n;\ \text{for all } n \tag{27}$$

where the matrices $\underline{A}_n$ and $\underline{H}_n$ in (27) have elements as indicated by the superscripts in (24) and (26).

The first term in (27) is a discrete convolution, so z-transform techniques can be used to express the solution. To this end we define

$$\begin{aligned} \underline{B}(f) &= \sum_{n} \underline{A}_n \exp\,[j2\pi nTf] \\ \underline{L}(f) &= \sum_{n} \underline{H}_n \exp\,[j2\pi nTf] \end{aligned} \tag{28}$$

and (27) is expressed as

$$\underline{B}(f)\underline{L}(f) + N_o\underline{L}(f) = \underline{B}(f) \tag{29}$$

The optimum filter is given by

$$\underline{H}_n = T \int_{-1/2T}^{1/2T} \underline{L}(f)\ \exp[-j2\pi\ nTf]df \tag{30}$$

where $\underline{L}(f) = [\underline{B}(f) + N_o\underline{I}]^{-1}\ \underline{B}(f)$.

The performance functional (15) is

$$J_o = \frac{N_o}{T} \operatorname{tr} \underline{H}_o = N_o \sum_i \int_{-1/2T}^{1/2T} L^{ii}(f)\, df \tag{31}$$

For the best time-invariant filter, we write the transfer function in terms of the power spectral density of the phase-randomized process as in (19). In some cases, it is convenient to express K(f) in terms of the $\{\underline{A}_n\}$ representation for the cyclostationary process. This will give

$$K(f) = \frac{1}{T} \sum_{\ell} \sum_{i} \sum_{j} \Lambda_{\ell}^{ij}\, \Phi_i(f)\, \Phi_j^*(f) \exp[-j\, 2\pi \ell T f] \tag{32}$$

III. EXAMPLES

Frequency-division Multiplexing (FDM).

A simple illustration of the application of the frequency-domain method is the situation where a set of uniformly-spaced, phase-locked, sinusoidal carriers are amplitude modulated with separate stationary signal processes. Let

$$x(t) = \sum_{i=o}^{M-1} x_i(t) \cos\left[2\pi \frac{(i_o+i)t}{T}\right];\ i_o > o \tag{33}$$

where the M processes. $\{x_i(t);\ i=0,1,\ldots,M-1\}$ are statistically independent, wide-sense stationary, and each bandlimited to the interval $|f| \leq 1/2T$. Then the representation (2) is simply

$$\begin{aligned} a_n(t) = a_{-n}(t) &= \tfrac{1}{2} x_n(t) \text{ for } i_o \leq n \leq i_o + M - 1 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{34}$$

Since the signals are independent, only the (+n,-n) pairs of the $\{a_n(t)\}$ are correlated and we have

$$\begin{aligned} r_{nm}(\tau) &= \tfrac{1}{4} k_i(\tau) \text{ for } |n|,\ |m| = i_o + i \\ &= 0 \text{ otherwise.} \end{aligned} \tag{35}$$

This particular form of the $\underline{R}(\cdot)$ matrix, means that only the inversion of a sequence of 2x2 matrices is needed for the solution of (14). The result is

$$G_{nn}(f) = \frac{\frac{1}{2} K_i(f)}{K_i(f) + 2N_o} = G_{n,-n}(f) \tag{36}$$

where $n = \pm(i_o + i)$ in (36). The +n and -n paths can be combined so that the overall filter has only M parallel paths; the i^{th} path with the bandlimited filter in (36) and multiplication at input and output by $2 \cos\left[2\pi \frac{(i_o + i)t}{T}\right]$.

From (16), the performance of the filter is given by

$$J_o = N_o \sum_n \int_{-\infty}^{\infty} G_{nn}(f)\, df = N_o \sum_{i=o}^{M-1} \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \frac{K_i(f)}{K_i(f) + 2\, N_o}\, df \tag{37}$$

whereas, for the time-invariant filter,

$$K(f) = \sum_n R_{nn}(f - \frac{n}{T}) = \frac{1}{4} \sum_{i=o}^{M-1} K_i(f - \frac{i_o+i}{T}) + K_i(f + \frac{i_o+i}{T})$$

which gives a performance,

$$J_1 = 2N_o \sum_{i=o}^{M-1} \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \frac{K_i(f)}{K_i(f) + 4\,N_o}\, df \qquad (38)$$

Comparing (37) and (38), we note that as signal-to-noise ratio increases, the time-invariant filter performance approaches 3 db. worse than the periodically-variable filter. Because of the similarity in (37) and (38) the two types of estimation can be compared on a channel-by-channel basis and large differences among the statistics of the $\{x_i(t)\}$ processes will not affect the comparison. This is not the case in the following example.

Time-division Multiplexing (TDM).

In this example, we again consider the multiplexing of M independent, stationary processes, each bandlimited to $|f| \leq \frac{1}{2T}$. The signals are sampled every T seconds to amplitude modulate pulses which are uniformly interleaved over the T second interval. The composite signal is

$$x(t) = \sum_{i=o}^{M-1} \sum_{n=-\infty}^{\infty} x_i(nT)\, \phi(t - nT - \frac{iT}{M}) \qquad (39)$$

A suitable pulse shape for the composite signal might be a rectangular, unit-amplitude pulse of width $\frac{T}{M}$ or a sinc pulse with zeros at multiples of T/M. The signal is already in the form of representation by (21) so the time-domain method of solution is natural. We let

$$\phi_i(t) = \left(\frac{M}{T}\right)^{\frac{1}{2}} \phi(t - \frac{iT}{M});\ i=0,1,\ldots,M-1 \qquad (40)$$

in order to satisfy (22). Then

$$A^{ij}_{n-m} = \frac{T}{M} E\,[x_i(nT)x_j(mT)] = \frac{T}{M}\, \delta_{ij}\; k_i(nT-mT) \qquad (41)$$

which is a diagonal matrix of order M because of the independence of the multiplexed signals. As a result, $\underline{H}_n$ in (27) is diagonal and a scalar equation for each element on the diagonal can be written.

$$\frac{T}{M} \sum_m k_i(nT-mT) H^{ii}_m + N_o H^{ii}_n = \frac{T}{M} k_i(nT) \qquad (42)$$

Solving for H^{ii}_o as indicated in (30), we derive the following expression for the performance functional.

$$J_o = \frac{N_o}{T} \sum_i H^{ii}_o = N_o \sum_{i=o}^{M-1} \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \frac{K_i(f)}{K_i(f) + MN_o}\, df \qquad (43)$$

For the time-invariant filter we can use (32), noting that $\underline{A}_\ell$ is diagonal, to obtain the power spectral density of the phase-randomized process.

$$K(f) = \frac{1}{MT} \sum_{m} \sum_{i} K_i(f-\frac{m}{T}) \; |\Phi_i(f)|^2 \qquad (44)$$

In particular, using a sinc pulse of bandwidth M/2T for $\phi(t)$, the performance of the time-invariant filter can be expressed as

$$J_1 = N_o \sum_{i=o}^{M-1} \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \frac{K_i(f) \qquad df}{\frac{1}{M} \sum_{j} K_j(f) + MN_o} \qquad (45)$$

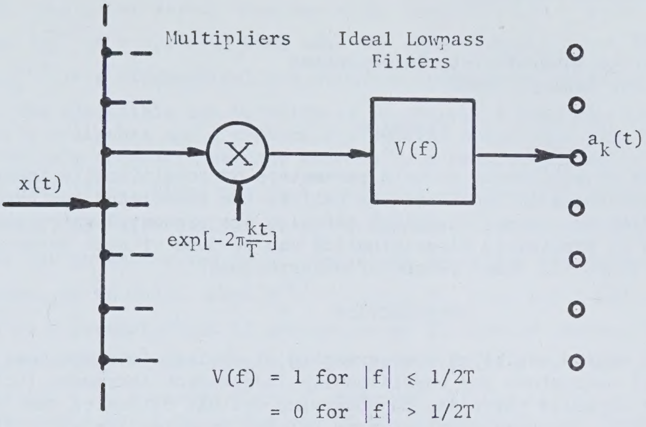
Comparing the time-invariant filter performance with that of the periodic filter, (45) and (43), we see that the difference narrows down to a term in the denominator of the integrand. For the time-invariant case, this term is the average power spectral density of all the processes multiplexed. One conclusion is that, if all the processes are statistically identical, then the two performance measures are identical. This is not surprising since, in the case where sinc pulses with bandwidth $\frac{M}{2T}$ are employed, the TDM signal would actually be wide-sense stationary. On the other hand, if there were large differences among the variances of the signal processes, the cyclostationary character becomes more pronounced and the periodically-variable filter can provide a substantial improvement. For example, suppose that all but one of the signal processes vanishes, then we have $J_1 \cong MJ_o$ in the low-noise case.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of this research by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-71-2111.

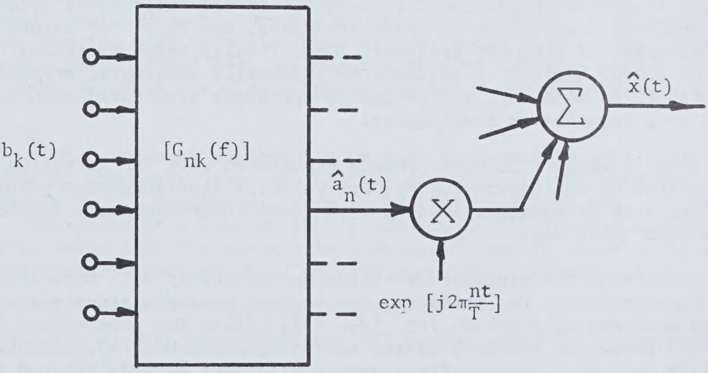
REFERENCES

1. W. R.Bennett,"Statistics of Regenerative Digital Transmission," Bell Sys. Tech. Jour., vol. 37, pp. 1501-42, November 1958.
2. A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, 1965.
3. E. G. Gladyshev, "Periodically and Almost Periodically Correlated Random Processes with Continuous Time Parameter," Theory Prob. Applications, pp. 173-177, 1963.
4. H. Hurd, "An Investigation of Periodically Correlated Stochastic Processes," Ph.D. Dissertation, Duke University, 1970.
5. H. Ogura, "Spectral Representation of a Periodic Nonstationary Random Process," IEEE Trans. Inform. Theory, vol. IT-17, pp. 143-149, March 1971.
6. L. E. Franks, Signal Theory, Prentice-Hall, 1969.


Multipliers
Ideal Lowpass Filters
x(t)
exp[-2π kt/T]
V(f)
a_k(t)

$$V(f) = 1 \text{ for } |f| \leq 1/2T$$
$$= 0 \text{ for } |f| > 1/2T$$

Fig. 1. Representation


b_k(t)
[G_nk(f)]
â_n(t)
exp [j2π nt/T]
x̂(t)

$$b_k(t) = \int_{-\infty}^{\infty} v(t-\tau)\exp[-j2\pi\frac{k\tau}{T}]z(\tau)d\tau = a_k(t) + n_k(t)$$

Fig. 2. Estimation

SOME SINGULAR ESTIMATION AND DETECTION PROBLEMS FOR CONDITIONALLY INDEPENDENT INCREMENT PROCESSES

P. A. FROST
Bell Telephone Laboratories, Incorporated
Whippany, New Jersey 07981

ABSTRACT

The problem of estimating certain parameters of conditionally independent increment processes is considered as well as the associated two hypothesis detection problem. Some "singular" examples are presented where parameter estimation or hypothesis discrimination can be done without error using an arbitrarily short (in time) record of observations.

INTRODUCTION

In [1] the author exhibited some examples of optimal (minimum mean square (m.m.s.)) estimators for conditionally independent increment (CII) processes. The emphasis there was on obtaining optimal estimates that are linear in the data. Here we consider some related examples for which the optimal estimates and related detection problems are 'singular'; that is, where the signal contained in a given realization of noisy observation data can be theoretically determined without error using a data record of arbitrarily short time duration. Hence the associated detection problem of discriminating between two possible signals from this class is also singular. Knowledge of singular estimation and detection problems, such as those discussed here, is essential if one is to avoid specifying ill-posed signal models.

We shall consider here in some detail three examples of singular estimation problems; a conditional gamma process, a conditional Cauchy process and a conditional Wiener process. Although it is not immediately evident from the original description of these processes, the first two belong to the same subclass of singular problems. Specifically these processes can be interpreted as the limit of a sequence of physically meaningful processes of the conditional Poisson type and estimation problems associated with each member of this sequence is nonsingular.

The term 'singular' is more closely associated with the detection literature than that of estimation theory, but we shall view detection problems as a special case of estimation and use the term interchangeably for both estimation and detection.

The conditions for singular detection are presently well understood for certain signal models. For instance, the case of gaussian signals may be considered substantially solved (cf. [2], [3]). Also for independent increment (II) processes the problem was solved by Skorokhod [4], although this work is less well known. The approach discussed here is related to Skorokhod's work, extending some of his results to certain CII processes. In some forthcoming companion papers the author developes some general formulas for estimation and detection in the context of CII signal models and also gives a representation of the likelihood ratio for CII processes which is analogous to Kailath's representation for random signals in additive gaussian white noise [5] (a model which is a special class of CII processes when signal and noise are independent).

CII PROCESSES

In this paper we may consider a CII process to be a doubly parameterized process $\{y_t^{(\mu)}, t \geq o, \mu \varepsilon U\}$ such that for each U contained in the admissible set $U, y_t^{(\mu)}$ is a stochastically continuous independent increment II process*. The admissible set U, which is in general a function space, is assumed to have a σ-algebra and a probability measure associated with it. In this paper we only consider the case where μ is a real number and U is a subset of the real line. Since for each $\mu \varepsilon U$, the process $y_t^{(\mu)}$ has independent increments, the distribution of $y_t^{(\mu)}$, which is necessarily an infinitely divisible law parameterized by μ, completely describes the process, provided we assume, as we shall, that $y_o^{(\mu)} \equiv o$, $\mu \varepsilon U$. Thus a CII process may be viewed as a parameterized II process or an II mixture process (cf. [7] for this terminology). The communication interpretation is that $y_t^{(\mu)}$ is a data process and μ is a random signal. Moreover, if the signal were known then the data would still be random and of the independent increment type. Since the formal derivative of an II process is a strict sense white noise, we are still dealing essentially with models of the signal contained in (not necessarily additive) white noise type. As noted above, the CII model includes the classical signal in independent additive white gaussian noise case which in the CII context is called a conditional Wiener process with drift. The CII model also includes the case of Poisson counting processes with stochastic intensity functions as considered by Snyder [8] and others. Actually any infinitely divisible distribution can be used to generate a CII process by giving it an admissible t and μ dependence.

It follows from the constructive way CII processes are defined that they have essentially the same sample function structure as II processes. Thus conditionally Wiener processes are the only members with continuous sample paths and the rest have step-type discontinuous paths. The derivative process for these 'jump' type processes therefore contains trains of impulses which suggests the circumstances where CII processes can be useful models.

In order to present our results we need to appeal to some basic facts concerning II processes, that are readily available in [9], and that carry over readily to CII processes. Here we briefly summarize some of these properties. The paths of a stochastically continuous CII process are either continuous in which case the process is also conditionally gaussian (a generalized conditional Wiener process) or have step-type discontinuities. A family of conditional Poisson counting processes $\{N_t^{(\mu)}(A), A \varepsilon R_\varepsilon = \{x : x|x| \geq \varepsilon, t \geq o\}\}$ can be associated with each jump type CII process with the property that $N_t^{(\mu)}(A)$ and $N_t^{(\mu)}(B)$ are conditionally independent if A and B are in R_ε and disjoint. $N_t^{(\mu)}(A)$ represents the number of jump changes in y that occurs in the interval [o,t) with size $\left(y_{s+o}^{(\mu)} - y_{s-o}^{(\mu)}\right) \varepsilon A$. Moreover if $y_t^{(\mu)}$ is a CII process then it can be represented in the form

---

* A more rigorous definition of a CII process must be formulated in terms of transition measures (cf. [6], p. 73).

$$y_t^{(\mu)} = \gamma_t^{(\mu)} + w_t^{(\mu)} + \int_{|\xi|\geq 1} \xi N_t^{(\mu)}(d\xi) + \int_{|\xi|<1} \xi\Big(N_t^{(\mu)}(d\xi) - \Gamma(t,d\xi,\mu)\Big) \qquad (1)$$

where $\gamma_t(\mu)$ is for each $\mu \varepsilon U$ a continuous nonrandom function, $w_t^{(\mu)}$ is a conditional Wiener process (see below) and $\Gamma(t,d\xi,\mu)$ is continuous in t and a measure on the Borel sets of R_ε for each $\mu \varepsilon U$. Also

$$E\Big[N_t^{(\mu)}(A)|\mu\Big] = \Gamma(t,A,\mu) \qquad (2)$$

and

$$E\Big[\exp\, i\phi y_t^{(\mu)}|\mu\Big] = \exp\left[i\phi\gamma_t^*(\mu) - \frac{\sigma^2}{2}\,\sigma_t^2(\mu) + \int \left(e^{i\phi\xi} - 1 - \frac{i\phi\xi}{1+\xi^2}\right)\Gamma(t,d\xi,\mu)\right] \qquad (3)$$

where $\sigma_t^2(\mu) \geq 0$ is continuous and nondecreasing in t for $\mu \varepsilon U$; $\sigma_t^2(\mu)$ is the conditional variance parameter of the C. Wiener process $w_t^{(\mu)}$. It is evident from (1) that if $y_t^{(\mu)}$ is a jump type CII process that $\{N_s^{(\mu)}(A), A \varepsilon R, s \leq t\}$ is equivalent to the record $\{y_s^{(\mu)}, 0 \leq s \leq t\}$ in the sense that the latter record can be obtained from the former. It is also clear from (1)-(3) that once $\Gamma(t,\cdot,\cdot)$, $\gamma_t^*(\cdot)$ and $\sigma_t^2(\cdot)$ are given the CII process is completely specified. In the balance of the paper we will, for notational simplicity write y_t and $N_t(A)$ in place of $y_t^{(\mu)}$ and $N_t^{(\mu)}(A)$. We now proceed to our examples.

A CONDITIONAL GAMMA PROCESS

Let y_t, $t \geq 0$ be a conditional 'Gamma' process with the gamma type conditional density

$$f_{y_t}(\lambda|\mu) = \frac{\alpha^{\mu t}}{\Gamma(\mu t)}\,\lambda^{\mu t-1}\, e^{-\alpha\lambda};\ \lambda > 0,\ \mu > 0,\ \alpha > 0 \qquad (4)$$

and associated conditional characteristic function

$$c_{y_t}(\phi|\mu) = E[\exp\, i\phi y_t|\mu] = 1/(1-i\phi/\alpha)^{\mu t} \qquad (5)$$

It is easily shown that (5) can also be expressed

$$c_{y_t}(\phi|\mu) = \exp\, \mu t \int_0^\infty (e^{i\phi\xi}-1)\,\frac{e^{-\alpha\xi}}{\xi}\, d\xi \qquad (6)$$

(In [1] we considered a conditional gamma process where α was the signal in which case no singular estimation problems occur.) If $\{N_t(A), A \varepsilon R_\varepsilon\}$ is the family of conditional Poisson counting processes associated with y_t, then from (6) it is evident that

$$E[N_t(A)|\mu] = \mu t \int_A \frac{e^{-\alpha\xi}}{\xi}\, d\xi \tag{7}$$

Now consider the following estimate, which is a natural one from the point of view of recording data from this process,

$$\hat{\mu}_{t,\varepsilon,n,m} \triangleq E\left[\mu|\Delta N_i(\Delta\lambda_j) = n_{ij};\ i = 0,1,\ldots,n-1, j = 0,1,\ldots,m-1\right] \tag{8}$$

where $0 = t_0 < t_1 < \ldots < t_n = t$, $\Delta t_i = t_{i+1} - t_i$; $0 < \varepsilon = \lambda_0 < \lambda_1 < \ldots < \lambda_{m-1} < \lambda_m = \infty$, $\Delta\lambda_i = (\lambda_i, \lambda_{i+1})$ and $\Delta N_i = N_{t_{i+1}} - N_{t_i}$. Following the method of [1] one finds directly that w.p.1.

$$\hat{\mu}_{t,\varepsilon,n,m} = \hat{\mu}_{t,\varepsilon} \triangleq E[\mu|N_\sigma((\varepsilon,\infty)), 0 \leq \sigma < t] = E[\mu|N_t((\varepsilon,\infty))]$$

$$= \frac{\int_0^\infty \lambda^{N_t((\varepsilon,\infty))+1}\, e^{-\lambda EI(\alpha\varepsilon)t}\, f_\mu(\lambda)d\lambda}{\int_0^\infty \lambda^{N_t((\varepsilon,\infty))}\, e^{-\lambda EI(\alpha\varepsilon)t}\, f_\mu(\lambda)d\lambda} \tag{9}$$

where $EI(\lambda) = \int_\lambda^\infty \frac{e^{-\xi}}{\xi}\, d\xi$. Thus in order to compute $\hat{\mu}_{t,\varepsilon}$ it is sufficient to simply record $N_t((\varepsilon,\infty))$; the number of jumps of size greater than a threshold ε. Note that from (8) it follows that

$$\hat{\mu}_{t,\varepsilon} \xrightarrow{\varepsilon\to 0} \hat{\mu}_t \triangleq E[\mu|y_\sigma, 0 \leq \sigma \leq t] \tag{10}$$

Now the minimum mean square linear estimate of μ in terms of $N_t((\varepsilon,\infty))$, denoted $\hat{\mu}^*_{t,\varepsilon}$, is easily determined to be

$$\hat{\mu}^*_{t,\varepsilon} = \frac{N_t((\varepsilon,\infty)) + \bar{\mu}\Big/\overline{(\mu-\bar{\mu})^2}}{\bar{\mu}\Big/\overline{(\bar{\mu}-\bar{\mu})^2} + t\int_\varepsilon^\infty e^{-\alpha\xi}\ \xi\ d\xi} \tag{11}$$

with associated mean square error

$$\overline{(\mu-\hat{\mu}^*_{t,\varepsilon})^2} = \frac{\bar{\mu}}{\bar{\mu}\Big/\overline{(\mu-\mu)^2} + t\int_\varepsilon^\infty e^{-\alpha\xi}/\xi\ d\xi} \tag{12}$$

Thus $\overline{(\mu-\hat{\mu}^{*}_{t,\varepsilon})^2} \xrightarrow{\varepsilon\to 0} 0$ for any $t > 0$. And since $\overline{(\mu-\hat{\mu}_{t,\varepsilon})^2} \le \overline{(\mu-\hat{\mu}^{*}_{t,\varepsilon})^2}$ it follows that $\hat{\mu}_{t,\varepsilon} \xrightarrow{\varepsilon\to 0} \mu$ w.p.1 for any $t > 0$ thereby demonstrating the singular nature of the estimate $\hat{\mu}_t$. However if we were using a conditional gamma process model of our data we would always have the constraint of being practically limited in our capability for measuring the small jumps of y_t. This would imply that we use $\hat{\mu}_{t,\varepsilon}$ as our estimate of μ which is quite well behaved (ε simply represents the resolving capability of our measuring instrument).

In connection with [1] it is interesting to note that formula (9) is essentially the same as that obtained in [1] for estimating the rate parameter of a Poisson step process from which we can conclude that $\hat{\mu}_{t,\varepsilon}$ will be linear in $N_t((\varepsilon,\infty))$ iff $f_\mu(\cdot)$ is a gamma density. If

$$f_\mu(\cdot) = \frac{\beta^q}{\Gamma(q)}\,\lambda^{q-1}\,e^{-\beta\lambda};\ \lambda > 0,\ \beta > 0,\ q > 0 \tag{13a}$$

then

$$\hat{\mu}_{t,\varepsilon} = \frac{N_t((\varepsilon,\infty)) + q}{\beta + tEI(\alpha\varepsilon)} \tag{13b}$$

and

$$\overline{(\mu-\hat{\mu}_{t,\varepsilon})^2} = \frac{q/\beta}{\beta + tEI(\alpha\varepsilon)} \tag{13c}$$

Note from (7) that $\{N_t((\varepsilon,\infty)), t \geq 0\}$ is a conditional Poisson counting process with conditional rate $\mu EI(\alpha\varepsilon)$ and as $\varepsilon \to 0$ we have an unbounded number of jumps for any $t \geq 0$. In [1] it was shown that for a Poisson counting process with a conditional rate $< \infty$, the rate parameter can still be determined without error as $t \to \infty$; by then we have observed an unbounded number of jumps. Thus in a certain sense this conditional gamma process case is quite analogous to the conditional Poisson case since in both cases a zero error is obtained in our estimate of the rates after an infinite number of jumps are observed.

A CONDITIONAL CAUCHY PROCESS

Suppose that y_t, $t \geq 0$ is a conditional Cauchy process with

$$f_{y_t}(\lambda|\mu) = \frac{1}{2\pi}\,\frac{\mu t}{\mu^2 t^2+\lambda^2}\,,\qquad \mu > 0 \tag{14a}$$

and corresponding characteristic function

$$C_{y_t}(\phi|\mu) = \exp -\mu t|\phi| \tag{14b}$$

It is easily shown that (14b) can be expressed

$$C_{y_t}(\phi|\mu) = \exp\left[\frac{\mu t}{\pi}\int_{-\infty}^{\infty}\left(e^{i\phi\xi} - 1 - \frac{i\phi\xi}{1+\xi^2}\right)\frac{d\xi}{\xi^2}\right] \tag{14c}$$

Thus if $\{N_t(A), A \in R_\varepsilon, \varepsilon > 0\}$ is the family of associated conditional Poisson counting processes, then

$$E[N_t(A)|\mu] = \frac{\mu t}{\pi} \int_A \frac{d\xi}{\xi^2} \tag{15}$$

Proceeding as in the previous example we obtain the result

$$\hat{\mu}_{t,\varepsilon} \triangleq E[\mu | N_\alpha(A), 0 \leq \sigma \leq t, A \in R_\varepsilon]$$

$$= E[\mu | N_t(R_\varepsilon)]$$

$$= \frac{\int_0^\infty \lambda^{N_t(R_\varepsilon)+1} e^{-2\lambda t/\pi\varepsilon} f_\mu(\lambda)d\lambda}{\int_0^\infty \lambda^{N_t(R_\varepsilon)} e^{-2\lambda t/\pi\varepsilon} f_\mu(\lambda)d\lambda} \tag{16}$$

Also the optimal linear estimate of μ in terms of $N_t(R_\varepsilon)$ is given by

$$\hat{\mu}^*_{t,\varepsilon} = \frac{N_t(R_\varepsilon) + \bar{\mu}\xi}{\xi + \frac{2t}{\pi\varepsilon}}, \quad \xi = \bar{\mu} \Big/ \overline{(\mu-\bar{\mu})^2} \tag{17a}$$

with mean squared error

$$\overline{(\mu-\hat{\mu}^*_{t,\varepsilon})^2} = \bar{\mu}/(\xi+2t/\pi\varepsilon) \tag{17b}$$

As in the previous example, the mean squared error (17b) goes to zero as ε tends to zero for any $t > 0$, thereby exhibiting a similar singularity.

A MORE GENERAL OBSERVATION

The above two processes, although quite different in various respects, exhibit similar singular properties. This similarity emerges more clearly when the underlying conditional Poisson counting processes are compared. In order to explore this similarity further, we note that in each case the family of conditional Poisson counting processes $\{N_t(A), A \in R_\varepsilon\}$ (fixed $\varepsilon > 0$) can be associated with a conditional compound Poisson process [10]. This process can be represented in the form

$$y_{t,\varepsilon} = \int_{|\xi|>\varepsilon} \xi N_t(d\xi) \tag{18a}$$

with

$$C_{y_{t,\varepsilon}}(\phi|\mu) = \exp \mu t k_\varepsilon \int_{|\xi|\geq\varepsilon} (e^{i\phi\xi}-1)\,dF_\varepsilon(\xi) \tag{18b}$$

where $F_\varepsilon(\cdot)$ is the step distribution. A comparison of (18b), (14c) and (6) shows that (18b) is obtained by simply exempting from the integrals (14c) and (6) an ε interval about the origin and subsequently normalizing the integrator. Thus the conditional gamma and Cauchy processes can each be considered as the limit of a particular sequence of conditional compound Poisson processes. In the conditional Gamma process case $k_\varepsilon = \int_\varepsilon^\infty (e^{-\alpha\xi}/\xi)d\xi$ and $F_\varepsilon(\alpha) = \int_\varepsilon^\alpha (e^{-\alpha\xi}/\xi)d\xi/k_\varepsilon$ and in the Cauchy case $k_\varepsilon = \int_{|\xi|>\varepsilon}^\infty \frac{d\xi}{\xi^2} = \frac{2}{\pi\varepsilon}$ and $F_\varepsilon(\alpha) = \frac{2}{\pi k_\varepsilon}\frac{1}{|\alpha|}$, $|\alpha| \geq \varepsilon$. We note that in each of these examples the 'signal' is proportional to the rate parameter of the compound Poisson process but the step distribution does not depend on μ. It is easily demonstrated, following the procedure of [1], that for any process of the form (18a-b)

$$E[\mu|y_{\sigma,\varepsilon}, 0 \leq \sigma \leq t] = E[\mu|N_t(R_\varepsilon)]$$

$$= \frac{\int_0^\infty \lambda^{N_t(R_\varepsilon)+1} e^{-\lambda t k_\varepsilon} f_\mu(\lambda)d\lambda}{\int_0^\infty \lambda^{N_t(R_\varepsilon)} e^{-tk_\varepsilon} f_\mu(\lambda)d\lambda} \tag{19}$$

The optimal linear estimate of μ in terms of $N_t(R_\varepsilon)$ is given by

$$\hat{\mu}^*_{t,\varepsilon} = \frac{N_t(R_\varepsilon) + \bar{\mu}\xi}{\xi + k_\varepsilon t} \tag{20a}$$

with mean squared error

$$\overline{(\mu-\hat{\mu}^*_{t,\varepsilon})^2} = \bar{\mu}/(\xi+tk_\varepsilon) \tag{20b}$$

where $\xi = \bar{\mu}/\overline{(\mu-\bar{\mu})^2}$. It also follows from (19), by using the technique of [1], that $E[\mu|N_t(R_\varepsilon)]$ is linear in $N_t(R_\varepsilon)$ iff has a gamma distribution as we have noted explicitely in the two examples above.

In summary, any conditional Poisson process which can be represented as the limit of a sequence of processes of the form (18a,b) for which $k_\varepsilon \to \infty$ as $\varepsilon \to 0$ will be singular in the sense defined in this paper. We also note that this circumstance can only occur when the jump measure $\int_A d_\xi\Gamma(t,\xi,\mu)$, (3), is unbounded for any set A that includes the origin. This implies that the process may be viewed suggestively as having an unbounded number of jumps, of arbitrarily small magnitude, in any finite time interval.

The sub-class of CII processes which can be obtained as the limit of a sequence of the form (18a,b) is rather small as compared with the entire class of CII processes, but it is not negligible as the examples given show.

In [11] we will derive a more general formula for estimating the signal μ when $\Gamma(t,\xi,\mu)$ does not factor as in the examples given here. The conditions for singularity in this general case can be related directly to Skorokhod's conditions [4] which apply for II processes.

A CONDITIONAL WIENER PROCESS

The final example of singular estimation we shall consider pertains to a conditional Wiener process with

$$C_{y_t}(\phi/\mu) = \exp -(\phi^2\mu^2 t/2) \tag{21}$$

It is a well known fact that the problem of discriminating between two Wiener processes having different variance parameters is singular. This example illustrates this fact in a slightly more general setting, but it is the approach used that we wish to draw attention to not any generality. We approach the question of singularity by considering the estimate of μ^2 from a sampled record of data $y_{t_0}, y_{t_1}, \ldots, y_{t_n}$ where $0 = t_0 < t_1 < \ldots < t_n = t$. Thus we compute

$$\begin{aligned} \widehat{\mu^2_n} &\overset{\Delta}{=} E[\mu^2 | y_{t_0}, \ldots, y_{t_n}] \\ &= E[\mu^2 | \Delta y_{t_0}, \ldots, \Delta y_{t_{n-1}}] \end{aligned} \tag{22}$$

where $\Delta y_{t_i} = y_{t_{i+1}} - y_{t_i}$. A simple application of Bayes rule and subsequent use of the conditional independence of the increments of y_t gives

$$\widehat{\mu^2_n} = \frac{\displaystyle\int_{-\infty}^{\infty} \lambda^2 |\lambda|^{-n} \exp\left(-\frac{v_n}{2\lambda^2}\right) f_\mu(\lambda) d\lambda}{\displaystyle\int_{-\infty}^{\infty} |\lambda|^{-n} \exp\left(-\frac{v_n}{2\lambda^2}\right) f_\mu(\lambda) d\lambda} \tag{23}$$

where

$$v_n = \sum_0^{n-1} \Delta y_{t_i}^2 \Big/ \Delta t_i$$

and $f_\mu(\cdot)$ is the probability density associated with μ. We note that $\widehat{\mu^2_n}$ can always be expressed as a function of v_n. Denoting the optimal linear estimate of μ^2 in terms of v_n by $\widehat{\mu^2_n}^{\,*}$ one easily obtains

$$\widehat{\mu^2_n}^{\,*} = \overline{\mu^2} + [v_n - \overline{\mu^2} n] \big/ (n+2+2\xi) \tag{24}$$

where $\xi = \overline{\mu^2}^{\,2} \big/ \overline{(\mu^2-\mu^2)^2}$. The associated error variance is

$$\overline{\left(\mu^2-\widehat{\mu^2_n}^{\,*}\right)^2} = 2\overline{\left(\mu^2-\overline{\mu^2}\right)^2}\;(1+\xi)/(n+2+\xi) \tag{25}$$

which obviously goes to zero as $n \to \infty$ for arbitrary t and $[0,t]$, arbitrary partitions and arbitrary $f_\mu(\cdot)$.

We note in connection with [1] that the optimal nonlinear estimate $\widehat{g(\mu)}_n$ will be linear in the data iff $g(\mu) = c_1\mu^2 + c_2$ (c_1 and c_2 arb. constants) and μ has a density of the form

$$f_\mu(\lambda) = k\sqrt{\frac{\beta}{2\pi}}\,\frac{1}{\lambda^{2p+2}}\,\exp - \beta/2\lambda^2 \tag{26}$$

where $p > 0$, $\beta >$ and $k = 1\cdot3\cdot5\cdots(2p-1)$. In this case, for n even

$$\widehat{\mu^2_n} = (v_n+\beta)/(n+2p-1) \tag{27a}$$

and

$$\overline{\left(\mu^2-\widehat{\mu^2_n}\right)^2} = 2\overline{\mu^4}\big/(n+2p) \tag{27b}$$

The singularity of this type of problem is often demonstrated by appealing to the quadratic variation (q.v.) of a Wiener process. If w_t is a Wiener process with parameter σ^2 then the q.v. $\sum_0^{n-1} (\Delta w_{t_i})^2 \xrightarrow{n\to\infty} \sigma^2 t$ w.p.1 and in q.m. It is easily shown that for a conditional Wiener process the q.v. converges in q.m. to $\mu^2 t$. The estimate

$$\widehat{\mu^2_n}^{\,+} = \frac{1}{t}\sum_0^{n-1} (\Delta y_{t_i})^2 \tag{28}$$

with error variance $\overline{\left(\mu^2-\widehat{\mu^2_n}^{\,+}\right)^2} = \dfrac{2\overline{\mu^4}}{t^2}\sum_0^{n-1} (\Delta t_i)^2$ also illustrates the desired singularity. It is interesting that the estimate (28), which doesn't use prior statistics, is inferior in a m.s.e. sense to $\widehat{\mu^2_n}^{\,*}$. In fact, it is inferior to the estimate $v_{n/n}$ which is the asymptotic form of $\hat{\mu}_n^{2*}$ unless equally spaced samples are used in which case they are equivalent.

REFERENCES

1. P. Frost, "Examples of Linear Solutions to Nonlinear Estimation Problems," Proc. of 5th Princeton Conference on Information Sciences and Systems, 1971.

2. T. Kailath, "Likelihood Ratios for Gaussian Processes," IEEE Trans. Inform. Theory IT-16, May 1970.

3. L. A. Shepp, "Radar-Nikodym derivatives of Gaussian Measures," Ann. Math. Statist. Vol. 37, Apr. 1966.

4. A. V. Skorokhod, Random Processes with Independent Increments, English Transl., Fed. Clearing House, AD 645 769, Aug. 1966.

5. T. Kailath, "A General Likelihood-Ratio Formula for Random Signals in Gaussian Noise," IEEE Trans. on Inform. Theory IT-15, May 1969.

6. J. Neveu, Mathematical Foundations of the Calculus of Probability, Holden Day, San Fran. 1965.

7. W. Feller, An Introduction to Probability Theory and Its Applications, Vol. II, J. Wiley, New York 1966.

8. D. Dnyder, "Filtering and Detection for Doubly-Stochastic Poisson Processes," To appear IEEE Trans. on Inf. Theory, Jan. 1972.

9. A. V. Skorokhod and I. I. Gikhman, Introduction to the Theory of Random Processes, W. B. Saunders Co., Philadelphia 1969.

10. E. Parzen, Stochastic Processes, Holden-Day, San Francisco 1962.

11. P. Frost, "Estimation and Detection for a Simple Class of Conditionally Independent Increment Processes," Proceedings of IEEE Decision and Control Conference, Miami, Dec. 1971.

NONRECURSIVE SMOOTHING FOR GAUSSIAN MODULATED POINT PROCESSES*

DONALD L. SNYDER
Biomedical Computer Laboratory
Washington University School of Medicine
St. Louis, Missouri 63110

ABSTRACT

We give a recursive equation for the time evolution of the posterior statistics of an information-bearing Markov process that modulates the intensity of an observed, self-exciting, doubly-stochastic point-process. The equation can be used to solve the filtering and detection problem for point process observations. We use it to derive a nonrecursive equation for the smoothed estimate of a Gaussian process that modulates the intensity.

INTRODUCTION

We associate with an observed point process on $T = (t_0, t_1]$ an integer valued counting process $\{N_t; t\varepsilon T\}$ and a counting record $N_{t_0,t} = \{(N_\sigma,\sigma);\ t_0<\sigma\leq t\}$ for $t\varepsilon T$. We assume that the following limits exist (a.s.):

$$\lim_{\Delta t\downarrow 0}(\Delta t)^{-1}\Pr(\Delta N_t=1|N_{t_0,t},\ \underline{x}_\sigma;\ t<\sigma\leq t+\Delta t)$$

$$= \lim_{\Delta t\downarrow 0}(\Delta t)^{-1}\Pr(\Delta N_t\geq 1|N_{t_0,t},\ \underline{x}_\sigma;\ t<\sigma\leq t+\Delta t) = \lambda_t(N_{t_0,t},\underline{x}_t)$$

where ΔN_t is the forward difference $N_{t+\Delta t}-N_t$ and $\underline{x}$ is a modulation process to be defined. This implies that with probability one points do not coincide and N has unit jumps. We take $N_{t_0}=0$ so that N_t is then the number of points in $(t_0,t]$.

In [1] we have described and given applications for doubly-stochastic Poisson processes for which λ_t is independent of $N_{t_0,t}$ and has the form $\lambda_t(N_{t_0,t},\ \underline{x}_t) = \mu_t(\underline{x}_t)$. Doubly-stochastic Markov birth processes are now included when λ_t depends on the count N_t and has the form $\lambda_t(N_{t_0,t},\ \underline{x}_t) = \mu_t(N_t,\ \underline{x}_t)$. More generally, we allow λ_t to depend on the entire counting record $N_{t_0,t}$. Rubin [2] has also examined this class of point processes but without the additional structure we impose on $\underline{x}$.

As in [1], let $\{\underline{x}_t;\ t\varepsilon T\}$ be an n-vector Markov process with initial state $\underline{x}_0$ at $t=t_0$.

---

*This work was supported by the National Institutes of Health Research Grant No. RR 00396 from the Division of Research Facilities and Resources.

EVOLUTION OF POSTERIOR STATISTICS

Let $c_t(\underline{v}|N_{t_0,t}) = E\{\exp[j<\underline{v}, \underline{x}_t>|N_{t_0,t}\}$ be the posterior characteristic function for $\underline{x}_t$ given an observed counting record. It can be shown that c satisfies the following recursive equation:

$$dc_t(\underline{v}|N_{t_0,t}) = E\{\exp[j<\underline{v},\underline{x}_t>]\, \Psi_t(\underline{v}|N_{t_0,t}, \underline{x}_t)|N_{t_0,t}\}dt \qquad (1)$$

$$+ E\{\exp[j<\underline{v},\underline{x}_t>](\lambda_t(N_{t_0,t}, \underline{x}_t) -\hat{\lambda}_t)|N_{t_0,t}\}\hat{\lambda}_t^{-1}\ (dN_t-\hat{\lambda}_t dt)$$

with the initial condition $c_{t_0}(\underline{v}|N_{t_0,t_0}) = E\{\exp[j<\underline{v}, \underline{x}_0>]\}$, where $\hat{\lambda}_t=E[\lambda_t(N_{t_0,t}, \underline{x}_t)|N_{t_0,t}]$ is the causal MMSE estimate of $\lambda_t(N_{t_0,t},\underline{x}_t)$ given $N_{t_0,t}$ and

$$\psi_t(\underline{v}|N_{t_0,t}, \underline{x}_t) = \lim_{\Delta t\downarrow 0}(\Delta t)^{-1}\ E\{(\exp[j<\underline{v},\Delta\underline{x}_t>] -1)|N_{t_0,t}, \underline{x}_t\}.$$

The proof of (1) is a trivial modification of that in [1] for the situation when λ_t is independent of $N_{t_0,t}$. Consequently, it will not be repeated here. Proceeding exactly as in [1], equation (1) can be used in the development of causal estimators for $\underline{x}_t$ and for solving the detection problem for counting processes of the type defined above. In fact, the equations of [1] remain unchanged except for incorporating the dependency of λ_t on $N_{t_0,t}$. Here, we use (1) to study the smoothing problem when $\underline{x}$ is replaced by a real valued Gaussian process.

NONRECURSIVE SMOOTHING

Let $\underline{x}_t = \underline{x}$, where $\underline{x}$ is a vector of random variables with prior p.d.f. $p_0(\underline{X})$. Then $\psi_t=0$ and inverse Fourier transformation of (1), use of the differential rule in [1], and integration shows

$$\ln p_{t_1}(\underline{X}|N_{t_0,t_1}) = \ln p_0(\underline{X}) - \int_{t_0}^{t_1}[\lambda_\sigma(N_{t_0,\sigma},\underline{X}) -\hat{\lambda}_\sigma]d\sigma \qquad (2)$$

$$+ \int_{t_0}^{t_1} \ln[\lambda_\sigma(N_{t_0,\sigma},\underline{X})/\hat{\lambda}_\sigma]dN_\sigma$$

where $p_t(\underline{X}|N_{t_0,t})$ is the posterior p.d.f. for $\underline{x}$ at time t.

Now consider the following smoothing problem. Let λ have the form $\lambda_t(N_{t_0,t}, x_t)$ where x_t is a real-valued Gaussian random-process with zero mean and covariance $K_{t,u}$. A counting record N_{t_0,t_1} is to be used to estimate the modulation waveform $\{x_\sigma;\ t_0<\sigma\leq t_1\}$. By using (2), we can demonstrate that the "maximum a posteriori probability estimate"

$\{\hat{x}_\sigma;\ t_0<\sigma\leq t_1\}$ of $\{x_\sigma;\ t_0<\sigma\leq t_1\}$ satisfies the following integral equation

$$\hat{x}_t = \int_{t_0}^{t_1} K_{t,\sigma}\{\partial \ln[\lambda_\sigma(N_{t_0,\sigma},\hat{x}_\sigma)]/\partial \hat{x}_\sigma\}\ [dN_\sigma - \lambda_\sigma(N_{t_0,\sigma},\ \hat{x}_\sigma)d\sigma],\ t_0<t\leq t_1 \qquad (3)$$

The procedure for establishing (3) is parallel to that used by Van Trees [3, Ch. 5] for an additive, white, Gaussian-noise model. We use (2) in connection with the coefficients in a truncated Karhunen-Loève representation for x. Let $x_t^{(K)} = \sum_{k=1}^{K} x_k \phi_k(t)$ be a K-term approximation to x_t on $(t_0,t_1]$ where $\{x_k\}$ are the Fourier coefficients of the representation and where $\{\phi_k\}$ are the normalized eigenfunctions of the integral equation

$$\mu_k \phi_k(t) = \int_{t_0}^{t_1} K_{t,u}\phi_k(u)du, \quad t_0<t\leq t_1$$

The coefficients $\{x_k\}$ are independent, zero-mean, Gaussian variables with covariances $\{\mu_k\}$, where μ_k is the eigenvalue associated with the eigenfunction ϕ_k.

Let $N^{(K)}$ denote the observed counting process when x is replaced by $x^{(K)}$ and let $\underline{x}$ be the K-vector of coefficients $\{x_k\}$. Then, from (2) we have

$$\partial \ln p_{t_1}(\underline{X}|N^{(K)}_{t_0,t_1})/\partial X_k = -\mu_k^{-1}X_k + \int_{t_0}^{t_1}\phi_k(\sigma)\{\partial \ln[\lambda_\sigma(N^{(K)}_{t_0,\sigma},x^{(K)}_\sigma)]/\partial x^{(K)}_\sigma\}\times$$
$$[dN^{(K)}_\sigma - \lambda_\sigma(N^{(K)}_{t_0,\sigma},x^{(K)}_\sigma)d\sigma] \qquad (4)$$

for k =1, 2,···, K. Let $\hat{x}^{(K)}_t = \sum_{k=1}^{K}\hat{x}_k\phi_k(t)$, $t_0<t\leq t_1$ where $\hat{x}_k$ is the "maximum a posteriori probability" estimate (posterior mode) of x_k given $N^{(K)}_{t_0,t_1}$. Then the left side of (4) is zero for $X_k = \hat{x}_k$ and we obtain:

$$\hat{x}^{(K)}_t = \int_{t_0}^{t} [\sum_{k=1}^{K}\mu_k\phi_k(t)\phi_k(\sigma)]\{\partial \ln[\lambda_\sigma(N^{(K)}_{t_0,\sigma},\ \hat{x}^{(K)}_\sigma)/\partial \hat{x}^{(K)}_\sigma\}\times$$
$$[dN^{(K)}_\sigma - \lambda_\sigma(N^{(K)}_{t_0,\sigma},\hat{x}^{(K)}_\sigma)d\sigma]$$

for $t_0<t\leq t_1$. Letting $K\to\infty$ and using Mercer's theorem then establishes (3).

The nonlinear integral equation in (3) cannot be solved explicitly in general. However, in parallel with nonlinear, nonrecursive smoothing for the additive white Gaussian noise observation model [3-6], it can form the point of departure for developing practical estimators. For instance, suppose $\lambda_t(N_{t_0,t},x_t) = A[1+m\cos(\omega t+x_t)]$ where A, m, and ω are known constants with A>0 and $|m|<1$. Then (3) implies that $\hat{x}$ can be generated by a form of noncausal tanlock loop [7]. Approximation procedures paralleling those described in [6] can presumably be used to realize an approximation to $\hat{x}$ that can be generated with a causal tanlock loop and a noncausal post-loop filter that can be realized with delay. In [1, Section 6.2.3] we specify the structure of this causal tanlock loop for estimating a Gaussian process x_t that has a finite-dimensional Gauss-Markov representation.

REFERENCES

1. D. L. Snyder, "Filtering and Detection for Doubly-Stochastic Poisson Processes," to appear in the IEEE Trans. on Information Theory; Jan., 1972.
2. I. Rubin, "Regular Point Processes and Their Detection," submitted for publication.
3. H. L. Van Trees, Detection, Estimation, and Modulation Theory: Part I, John Wiley and Sons, Inc., New York; 1968.
4. D. C. Youla, "The Use of Maximum Likelihood in Estimating Continuously Modulated Intelligence Which Has Been Corrupted by Noise," IRE Trans. Inform. Theory, IT-3, 90-105, March, 1954.
5. H. L. Van Trees and C. Boardman, "Optimum Angle Modulation," IEEE Trans. on Comm. Tech., COM-13, No. 4, pp. 452-469; Dec., 1965.
6. H. L. Van Trees, Detection, Estimation, and Modulation Theory: Part II, John Wiley and Sons, Inc., New York; 1971.
7. L. M. Robinson, "Tanlock: a phase-lock loop of extended tracking capability," Proc. Nat'l. Winter Conf. on Military Electronics; 1962.

JOINT DETECTION, ESTIMATION AND SYSTEM IDENTIFICATION[1]

D.G. LAINIOTIS
Department of Electrical Engineering and Electronics Research Center
The University of Texas at Austin
Austin, Texas 78712

ABSTRACT

Recent results of Middleton and Esposito [1968] and Lainiotis [1969] on single-shot joint detection-estimation for discrete data are extended to the single-shot continuous data case and generalized to joint Bayesian detection-estimation-system identification. Moreover, previous results were generalized to the case of causal estimator. Specifically, it is shown that the above problem constitutes a class of nonlinear mse estimation problems, with the attendant difficulties in realizing the optimal nonlinear estimators. However, by utilizing the adaptive approach closed form integral expressions are given. These are given in terms of the generalized likelihood ratio $\Lambda(t)$, which is a sufficient statistic for Bayes-optimal compound detection. The latter in turn is specified by a continuum (for continuous θ) θ-conditional likelihood ratious $\Lambda(t/\theta)$ each of which is the LR for testing for the model specified by the parameter value θ. The latter LR's are, moreover, given in terms of optimal mse causal estimators. In essence then, it has been shown that system identification is equivalent to multihypothesis testing, with a continuum or finite sequence of hypotheses, respectively for continuous or finite discrete range of θ.

I. INTRODUCTION

The problem considered here is the single-shot, compound detection-estimation problem specified by the following equations:

$$z(t) = s(t) + v(t) = \beta y(t) + v(t) \tag{1}$$

where $\{z(t)\}$ is the m-vector observable random process, $v(t)$ is the white gaussian, zero-mean observation noise random process with covariance matrix $R(t)$, and β is the so-called indicator variable which takes values 1 or 0 depending on whether H_1 (signal present) or H_0 (signal absent) is true, with a-priori probability p_1, p_0, respectively, $\left(\sum_0^1 p_i = 1 \right)$.

Moreover, the signal random process $\{y(t)\}$ is assumed to be "adequately" modeled by the state-variable model

$$y(t) = h(x(t), t; \theta) \tag{2a}$$

and

$$\frac{dx(t)}{dt} = f(x(t), t; \theta) + g(x(t), t; \theta)\, u(t) \tag{2b}$$

where $\{x(t)\}$ is the n-vector signal "state" random process, and $\{u(t)\}$ is a zero-mean q-vector white gaussian process, independent of $\{v(t)\}$, whose covariance matrix is the identity matrix. The functionals $h(\cdot)$, $f(\cdot)$, and $g(\cdot)$

are time-varying nonlinear functionals of the state vector $x(t)$, and of θ. The s-vector parameter θ is time-invariant and, if known, specifies the above model completely. However, in the compound problem considered, θ is unknown and following the Bayesian approach, it is considered a random variable with known or assumed a-priori density $p(\theta)$.

The initial state-vector $x(t_o)$ is independent of $\{v(t)\}$ and $\{u(t)\}$ for $t \geq t_o$ and has a known θ-conditional gaussian a-priori density, denoted by $p(x(t_o)/t_o, \theta)$ with means $\hat{x}(t_o/t_o, \theta)$ and covariance matrix $P(t_o/t_o, \theta)$.

Given the continuous record $\nu_t = \{z(\tau); \tau \in (t_o, t)$ for the current baud, for which a decision is to be made, the Bayes-risk minimizing decision rule as well as the minimum mean-square estimates of the signal random process $s(t)$ and the unknown parameter θ, are to be derived.

Middleton and Esposito [1968], considered the single-shot joint detection estimation problem for discrete data. Specifically, for both the strong and weak-coupling case and for quadratic-cost function (mse) estimation of $s(t)$, they obtained the following fundamental relationship

$$\hat{s}(t/\nu_n) = \frac{\Lambda_n}{1+\Lambda_n} \hat{s}(t/\nu_n, \beta = 1) \tag{3}$$

where $\hat{s}(t/t) \equiv E[s(t)/\nu_n]$, $\hat{s}(t/\nu_n, \beta = 1) \equiv E[s(t)/\nu_n, \beta = 1]$ and Λ_n is the generalized likelihood ratio, Middleton and Esposito [1968], defined as

$$\Lambda_n \equiv \frac{\rho \int p(\nu_n/\beta = 1, \theta) p(\theta) d\theta}{p(\nu_n/\beta = 0)}, \quad \rho \equiv \frac{p_1}{p_o} \tag{4}$$

where $\nu_n = \{z(t_1), z(t_2), \ldots, z(t_n); t_i \in (t_o, t)\}$ and $p(\nu_n/\beta = 1, \theta)$ is the θ-parameter conditional density of ν_n under H_1, and $p(\nu_n/\beta = 0)$ is the probability density of ν_n under H_0. Note θ is connected to the signal only. No other restrictions were imposed on the model.

Lainiotis [1969], moreover, recently established the fundamental fact that detection is indeed mean-square estimation of the indicator variable β, for the model of Eq. 1. Specifically, he established that the Bayes-optimal decision procedure is: decide H_1 or H_0 depending on whether $\hat{\beta}(\lambda_n) \gtrless c_o$, respectively, where c_o depends on the a-priori costs for each decision, and $\hat{\beta}(\nu_n) \equiv E[\beta/\nu_n]$. Moreover, Lainiotis [1969], showed that

$$\hat{\beta}(\nu_n) = p(\beta = 1/\nu_n) = \frac{\Lambda_n}{1 + \Lambda_n} \tag{5}$$

Note that Eq. (5) above is a special case, Esposito [1969], of Middleton and Esposito's formula, since $\hat{\beta}(\nu_n/\beta = 1)$ is obviously one.

In this paper, the above results of Middleton and Esposito [1968] and Lainiotis [1969] are extended to the continuous data case and generalized to joint detection-estimation-system identification, as well as to the case of causal mse estimators.

II. OPTIMAL JOINT DETECTION, ESTIMATION AND SYSTEM IDENTIFICATION

The problem stated in the introduction constitutes joint detection (i.e. β) signal estimation (i.e. $s(t)$), and system identification (i.e. θ). It may be shown to be simply a nonlinear mean-square estimation problem. To see this, augment the state-vector $x(t)$ with θ and β, so that the augmented state-vector $x_a(t) \equiv [x^T(t) : \theta^T : \beta]^T$. Then the model defining equations become

$$\frac{dx_a(t)}{dt} = f_a(x_a(t), t) + g_a(x_a(t), t)\, u(t) \tag{6a}$$

$$z(t) = h_a(x_a(t), t) + v(t) \tag{6b}$$

where $f_a(x_a(t), t) \equiv [f^T(x(t), t; \theta) : 0 : 0]^T$, $g_a(x_a(t), t) \equiv [g^T(x(t), t; \theta): 0:0]^T$ and $h_a(x_a(t), t) \equiv \beta h(x(t), t; \theta)$. It is apparent from their definitions that $f_a(\cdot)$, $g_a(\cdot)$, and $h_a(\cdot)$ are, in general, nonlinear functionals of $x_a(t)$.

In view of the discussion in the introduction it is readily seen that the optimal mean-square estimate (mse) of $x_a(t)$ given ν_t, denoted $\hat{x}_a(t/t) \equiv [\hat{x}^T(t/t): \hat{\theta}^T(t): \hat{\beta}(t)]^T$, contains all the quantities for the solution of the joint minimum Bayes-risk detection and mse estimation and system identification problem. Thus, the above problem is equivalent to mse nonlinear estimation, with the attendant difficulties in realizing nonlinear estimators, Jazwinski [1970].

In this paper, by utilizing the adaptive approach, closed form formulas are obtained. The desired adaptive realization is obtained by considering both θ and β as an augmented parameter vector $\alpha = [\theta^T : \beta]^T$. The adaptive realization is given in the following.

Theorem I: (Partition Theorem)

$$\hat{\beta}(t) = \frac{\rho \int \Lambda(t/\theta)\, p(\theta) d\theta}{1 + \rho \int \Lambda(t/\theta) p(\theta) d\theta} \tag{7a}$$

$$\hat{s}(t/t) = \frac{\rho\int\Lambda(t/\theta)p(\theta)d\theta}{1 + \rho\int\Lambda(t/\theta)p(\theta)d\theta} \cdot \hat{s}(t/t\,,\ \beta = 1) \tag{7b}$$

$$\hat{\theta}(t) = \hat{\theta}(t_o)p(\beta = 0/t) + \frac{\rho\int\theta\Lambda(t/\theta)p(\theta)d\theta}{1 + \rho\int\Lambda(t/\theta)p(\theta)d\theta} \tag{7c}$$

where

$$\hat{\beta}(t) \equiv E[\beta/t] = p(\beta = 1/t) = 1 - p(\beta = 0/t).$$

$$\hat{s}(t/t) \equiv E[s(t)/t]$$

$$\hat{s}_1(t/t) \equiv E[s(t)/t,\ \beta = 1]$$

$$\hat{\theta}(t) \equiv E[\theta/t]$$

and where

$$p(\theta/t) = \frac{1 + \rho\Lambda(t/\theta)}{1 + \rho\int\Lambda(t/\theta)p(\theta)d\theta}\ p(\theta) \tag{7d}$$

$$\hat{\theta}(t_o) = E[\theta/t_o] = \text{the a-priori mean}$$

and

$$\Lambda(t/\theta) \equiv \exp\left\{ \int_{t_o}^{t} \hat{h}_1^T(\sigma/\sigma,\ \theta)R^{-1}(\sigma)z(\sigma)d\sigma - \tfrac{1}{2}\int_{t_o}^{t} \|\hat{h}_1(\sigma/\sigma,\ \theta)\|^2_{R^{-1}(\sigma)}\, d\sigma \right\}. \tag{7e}$$

where $\hat{h}_1(\sigma/\sigma,\ \theta) \equiv E[h(x(\sigma),\ \sigma;\ \theta)/\sigma,\ \theta,\ \beta = 1]$.

Proof: The proof is given in Lainiotis [1971].

The following remarks on the above results are pertinent:

It is seen from Eq. (7a) that the statistic sufficient for Bayes-optimal detection is $\Lambda(t) \equiv \rho\int\Lambda(t/\theta)p(\theta)d\theta$. $\Lambda(t)$ is the continuous data generalized likelihood ratio for compound detection. In terms of $\Lambda(t)$, Eq. (7a) is the same as the one obtained by Lainiotis [1969] for discrete data.

Moreover, we note that the generalized likelihood ratio is specified by a continuum of θ - conditional likelihood ratios (LR) $\Lambda(t/\theta)$, each of which is the LR for the detection problem

$$H_1:\ z(t) = h(x(t),\ t;\ \theta) + v(t)$$

$$H_0:\ z(t) = v(t)$$

for a specified admissible value (admissible in the sense of $p(\theta)$). In other words, $\Lambda(t/\theta)$ is the LR for testing whether the signal generated by the model specified by parameter value θ is present. This leads us to the conclusion that, essentially, system identification is equivalent to multihypothesis testing with a continuum of hypotheses (for continuous θ) corresponding to each possible model indexed by θ.

In addition, we note from Eq. (7e) that $\Lambda(t/\theta)$ is in the canonical estimator-correlator form of Kailath [1969], which is particularly well-suited to interpretation and approximation. However, it must be pointed out that the generalized likelihood ratio for continuous data $\Lambda(t)$ is applicable to Bayes-optimal compound detection and as such is more general than Kailath's [1969] LR, the latter being applicable to classical hypothesis testing without consideration of prior probabilities.

Equation (7b) above is the continuous data version of Middleton and Esposito's [1969] result. We note that is has similar form and interpretation as their result, with, however, a very significant difference. Namely, Middleton and Esposito's [1969] results were valid, essentially, for noncausal estimators since in the discrete approach used in their paper the waveform is estimated as a whole given all the data. In contrast the results given herein, generalize Middleton and Esposito's [1969] results not only to the continuous case but also to the case of causal estimators.

Using Eq. (7a), $\hat{s}(t/t)$ may be given the interesting form

$$\hat{s}(t/t) = \hat{\beta}(t)\hat{s}_1(t/t). \tag{8}$$

Similarly, for $\hat{\theta}(t)$, we have

$$\hat{\theta}(t) = \hat{\theta}(t_o)[1 - \hat{\beta}(t)] + \frac{\rho\int\theta\Lambda(t/\theta)p(\theta)d\theta}{1 + \rho\int\Lambda(t/\theta)p(\theta)d\theta} \tag{9}$$

Eqs. (8-9) establish once more the fact that optimal estimation of $s(t)$ and θ requires nonlinear processing of the data. This is true even in the case of linear models, namely, when $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are linear functionals of the state-vector $x(t)$.

In the following corollary, expressions for the mse estimates of $h(\cdot)$ and $s(t)$ as well as the corresponding conditional error-covariance matrices are given.

Corollary I:

$$\hat{s}(t/t) = \hat{\beta}(t)\hat{h}_1(t/t) \tag{10a}$$

$$P_s(t/t) = \hat{\beta}(t)P_{h_1}(t/t) \tag{10b}$$

where $\hat{h}_1(t/t) \equiv E[h(x(t), t, \theta)/t, \beta = 1]$,

$$P_s(t/t) \equiv E\{[s(t) - \hat{s}(t/t)][s(t) - \hat{s}(t/t)]^T/t\}$$

and $P_{h_1}(t/t) \equiv E\{[h(x(t), t, \theta) - \hat{h}_1(t/t)][h(x(t), t, \theta) - \hat{h}_1(t/t)]^T/t, \beta = 1\}$

$$\hat{h}_1(t/t) = \int \hat{h}_1(t/t, \theta) p_1(\theta/t)\, d\theta \tag{10c}$$

where $\hat{h}_1(t/t, \theta)$ was defined previously and $p_1(\theta/t)$ is given by

$$p_1(\theta/t) \equiv p(\theta/t, \beta = 1) = \frac{\Lambda(t/\theta)}{\int \Lambda(t/\theta p(\theta) d\theta}\; p(\theta) \tag{11}$$

where $p(\theta)$ is the a-priori density of θ; ρ and $\Lambda(t/\theta)$ were defined previously.

The conditional error-covariance matrix $P_{h_1}(t/t)$ is given by

$$P_{h_1}(t/t) = \int \{P_{h_1}(t/t, \theta) + [\hat{h}_1(t/t, \theta) - \hat{h}_1(t/t)][\hat{h}_1(t/t, \theta) - \hat{h}_1(t/t)]^T\} p_1(\theta/t) d\theta \tag{12}$$

where $P_{h_1}(t/t, \theta)$ and $\hat{h}_1(t/t, \theta)$ are given, Jazwinski [1970], as the solution of coupled partial differential equations, whose solution is possible only for special cases, e.g., for linear models.

Proof: The proof is given in Lainiotis [1971].

At this point we note that the realization of the nonlinear estimator and the associated computational requirements depend essentially on the range of admissible values for θ, namely whether the range of θ is discrete (finite) or continuous. In most applications the range is continuous resulting in excessive computational requirements. One approach for alleviating the problem is to quantize the θ - space. Moreover, such quantization is reasonable in view of the fact that quantization occurs naturally in any physical realization of a system. In any case, either because θ is discrete or because of quantization of a continuous range, $p(\theta) \simeq \sum_{i=1}^{N} p(\theta_i)\, \delta(\theta - \theta_i)$. For such $p(\theta)$ all integrations with respect to θ in theorem I and corollary I become sums of N terms, each term corresponding to a particular value of θ, θ_i, $i = 1, 2, \ldots, N$.

It was shown earlier that system identification constitutes in essence a hypothesis testing problem with a continuum of hypotheses. Parameter space quantization, however, has reduced the problem to one with a finite set of hypotheses. In this context, system identification constitutes a sequence of hypotheses testing problems each corresponding to testing for the model indexed by parameter value θ_i, $i = 1, 2, \ldots, N$.

Applying Itô's, Jazwinski [1970], differentiation rule to Eqs. (7a, 7d, 11) we obtain stochastic differential equations for the temporal evolution of $\hat{\beta}(t)$, $p(\theta/t)$ and $p_1(\theta/t)$. These are given in the following corollary.

Corollary II:

$$d\hat{\beta}(t) = \hat{\beta}(t)\ [1 - \hat{\beta}(t)]\ \hat{h}_1^T(t/t)R^{-1}(t)\ [dz(t) - \hat{\beta}(t)\ \hat{h}_1(t/t)\ dt]$$

with initial condition $\hat{\beta}(t_o) = p_1$, and

$$dp_1(\theta/t) = p_1)\theta/t)\ [\hat{h}_1(t/t,\ \theta) - \hat{h}_1(t/t)]R^{-1}(t)\ [dz(t) - \hat{\beta}(t)\hat{h}_1(t/t)dt]$$

with initial condition $p_1(\theta/t_o) = p(\theta)$

$$dp(\theta/t) = \hat{\beta}(t)\ [p_1(\theta/t)\hat{h}_1^T(t/t,\ \theta) - p(\theta/t)\hat{h}_1(t/t)]R^{-1}(t)\ [dz(t) - \hat{\beta}(t)\hat{h}_1(t/t)dt]$$

with initial condition $p(\theta/t_o) = p(\theta)$.

Proof: The proof consists of straightforward application of Itô's differentation rule to Eqs. (7a, 7d, 11), and as such is omitted.

Again, it must be emphasized that algorithms for the evaluation of $\hat{h}_1(t/t,\ \theta)$ are needed. Such algorithms are possible for special cases only, such as the case of linear models. This case is treated in the following section.

We note that both integral expressions as well as stochastic differential equations for $\hat{\beta}(t)$, $p_1(\theta/t)$ and $p(\theta/t)$ were given above. However, in the opinion of this author, the integral expressions are far more valuable from a practical standpoint, namely, implementation as well as from an interpretation point of view. These opinions are based on the fact that the stochastic differential equations are nonlinear and coupled, and their solution is in general not easily forthcoming. Moreover, approximate solutions of these equations are ad-hoc and the effect of the approximations made cannot easily be assessed. In contrast, note that the effects of approximations made in evaluating the integral expressions, such as the quantization of the θ - range, can be easily assessed and improved upon--namely finer quantization--if desired.

III. SPECIAL CASE: LINEAR MODELS

In this section, the case of linear dynamic models is considered. Namely, the signal random process is given by $y(t) = H(t,\ \theta)x(t)$ (13a)

and $$\frac{dx(t)}{dt} = F(t,\ \theta)x(t) + G(t,\ \theta)u(t) \qquad (13b)$$

where $\{y(t)\}$, $\{u(t)\}$, and $\{x(t)\}$ were previously defined. The results for this case are summarized in the following corollary.

Corollary III:

$$\hat{s}(t/t) = \hat{\beta}(t)\hat{y}_1(t/t) \qquad (14a)$$

$$P_s(t/t) = \hat{\beta}(t)[\int\{H(t,\theta)P_1(t/t,\theta)H^T(t,\theta) + [H(t,\theta)\hat{x}_1(t/t,\theta) - \hat{y}_1(t/t)] \cdot [H(t,\theta)\hat{x}_1(t/t,\theta) - \hat{y}_1(t/t)]^T\}p_1(\theta/t)\,d\theta] \tag{14b}$$

where $\hat{y}_1(t/t) = \int H(t,\theta)\hat{x}_1(t/t,\theta)p_1(\theta/t)d\theta$ and $\hat{\beta}(t)$ and $p_1(\theta/t)$ are given by the stochastic differential equations:

$$d\hat{\beta}(t) = \hat{\beta}(t)[1 - \hat{\beta}(t)]\hat{y}_1^T(t/t)R^{-1}(t)\,[dz(t) - \hat{\beta}(t)\hat{y}_1(t/t)dt] \tag{14c}$$

with initial condition $\hat{\beta}(t_o) = p_1$, and

$$dp_1(\theta/t) = p_1(\theta/t)[H(t,\theta)\hat{x}_1(t/t,\theta) - \hat{y}_1(t/t)]R^{-1}(t)[dz(t) - \hat{y}_1(t/t)dt] \tag{14d}$$

with initial condition $p_1(\theta/t_o) = p(\theta)$.

$$dp(\theta/t) = \hat{\beta}(t)[p_1(\theta/t)H(t,\theta)\hat{x}_1(t/t,\theta) - p(\theta/t)\hat{y}_1(t/t)]R^{-1}(t)[dz(t) - \hat{\beta}(y)\hat{y}_1(t/t)dt]$$

with initial condition $p(\theta/t_o) = p(\theta)$. (14e)

Integral expressions for $\hat{\beta}(t)$ and $p_1(\theta/t)$ are given by Eqs. (7a) and (11), where now $\Lambda(t/\theta)$ takes the form

$$\Lambda(t/\theta) \equiv \exp\{\int_{t_o}^{t} \hat{x}_1^T(\sigma/\sigma,\theta)H^T(\sigma,\theta)R^{-1}(\sigma)z(\sigma)d\sigma - \tfrac{1}{2}\int_{t_o}^{t} \|H(\sigma,\theta)\hat{x}_1(\sigma/\sigma,\theta)\|^2_{R^{-1}(\sigma)}\,d\sigma\} \tag{15}$$

The signal and model conditional estimate $\hat{x}_1(t/t,\theta)$ and the corresponding error-covariance matrix $P_1(t/t,\theta)$ are now given by the well-known Kalman-Bucy, Jazwinski [1970], equations:

$$d\hat{x}_1(t/t,\theta) = F(t,\theta)\hat{x}_1(t/t,\theta) + P_1(t/t,\theta)H^T(t,\theta)R^{-1}(t)[dz(t)-H(t,\theta)\hat{x}_1(t/t,\theta)] \tag{16a}$$

with initial condition $\hat{x}_1(t/t,\theta) = \hat{x}(t_o/t_o)$, and

$$\frac{dP_1(t/t)}{dt} = F(t,\theta)P_1(t/t,\theta)+P_1(t/t,\theta)F^T(t,\theta)+G(t,\theta)G^T(t,\theta) - P_1(t/t,\theta)H^T(t,\theta) R^{-1}(t)H(t,\theta)P_1(t/t,\theta) \tag{16b}$$

with initial condition $P_1(t_o/t_o) = P(t_o/t_o)$.

Proof: The proof is based on simple application of Theorem I and Corollary I to the linear model given in Eqs. (13), and as such is omitted.

IV. SUBOPTIMAL NONLINEAR ESTIMATION ALGORITHM

It was shown in section II, that the problem of joint detection-estimation-system identification constitutes a nonlinear estimation problem. But, as was pointed out earlier, the optimal nonlinear filter is specified, in general, by an infinite set of coupled stochastic partial differentialequations. Such specification is not useful since it is in general unrealizable. This problem was partially alleviated by using an adaptive approach with quantization of the unknown parameter space. In this section an approximate nonlinear estimation algorithm is used that does not require parameter space quantization. This is based on the so-called relinearized Kalman-Bucy filter, Licht [1970].

It is widely known that the simplest method of suboptimal nonlinear filtering is that of Kalman-Bucy filtering about a nominal trajectory. Licht [1970] demonstrates that if the optimal estimate instead of fixed a-priori trajectory is chosen as the nominal trajectory, the resulting approximate relinearized filter is much more effective than the usual linearized filter about constant nominal trajectory.

The relinearized Kalman-Bucy filter for the model defined by (6a) and (6b) is given by the following set of differential equations:

$$\frac{d\hat{x}_a(t/t)}{dt} = f_a(\hat{x}_a(t/t),t) + P(t/t)\nabla h_a^T(\hat{x}_a(t/t),t)R^{-1}(t)[z(t)-h_a(\hat{x}_a(t/t),t)] \tag{17a}$$

$$\begin{aligned}\frac{dP(t/t)}{dt} &= \nabla f_a(\hat{x}_a(t/t),t)P(t/t) + P(t/t)\nabla f_a^T(\hat{x}_a(t/t),t) + g_a(\hat{x}_a(t/t),t)\cdot g_a^T(\hat{x}_a(t/t),t)\\ &\quad - P(t/t)\nabla h_a^T(\hat{x}_a(t/t),t)R^{-1}(t)\nabla h_a(\hat{x}_a(t/t),t)P(t/t)\end{aligned} \tag{17b}$$

where

$\hat{x}_a(t/t) = E[x_a(t)/t]$

$P(t/t) = E[[x_a(t) - \hat{x}_a(t/t)]\,[x_a(t) - \hat{x}_a(t/t)]^T]$ with initial conditions

$\hat{x}_a(t_o/t_o) = E[x_a(t_o)]$

$P(t_o/t_o) = E[[x_a(t_o)-\hat{x}_a(t_o/t_o)][x_a(t_o)-\hat{x}_a(t_o/t_o)]^T]$ where $\nabla h_a(x_a(t/t),t)$

and $\nabla f_a(x_a(t/t),t)$ are Jacobian matrices of $h_a(x_a(t/t),t)$ and $f_a(x_a(t/t),t)$ respectively.

In order to illustrate the implementation of the relinearized filter, the linear model in section III is considered. The results are given in the following corollary.

Corollary IV:

$$\begin{bmatrix} d\hat{x}(t/t) \\ d\hat{\theta}(t) \\ d\hat{\beta}(t) \end{bmatrix} = \begin{bmatrix} F(t,\hat{\theta}(t))\hat{x}(t/t) \\ 0 \\ 0 \end{bmatrix} dt + P(t/t)\begin{bmatrix} \hat{\beta}(t)H^T(t,\hat{\theta}(t)0 \\ \hat{\beta}(t)\hat{x}^T(t/t)\nabla H^T(t,\hat{\theta}(t)) \\ \hat{x}^T(t/t)H^T(t,\hat{\theta}(t)) \end{bmatrix}\cdot$$

$$\cdot R^{-1}(t)[dz(t) - \hat{\beta}(t)H(t,\hat{\theta}(t))\hat{x}(t/t)dt] \tag{18a}$$

$$dP(t/t) = \begin{bmatrix} F(t,\hat{\theta}(t)) & F(t,\hat{\theta}(t))\hat{x}(t/t) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} P(t/t)dt +$$

$$P(t/t) \begin{bmatrix} F^T(t,\hat{\theta}(t)) & 0 & 0 \\ \hat{x}^T(t/t)\nabla F^T(t,\hat{\theta}(t)) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} dt + \begin{bmatrix} G(t,\hat{\theta}(t))G^T(t,\hat{\theta}(t)) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} dt$$

$$- P(t/t) \begin{bmatrix} \hat{\beta}(t)H^T(t,\hat{\theta}(t)) \\ \hat{\beta}(t)\hat{x}^T(t/t)\ \nabla H^T(t,\hat{\theta}(t)) \\ \hat{x}^T(t/t)\ H^T(t,\hat{\theta}(t)) \end{bmatrix} R^{-1}(t) \begin{bmatrix} \hat{\beta}(t)H^T(t,\hat{\theta}(t)) \\ \hat{\beta}(t)\hat{x}^T(t/t)\nabla H^T(t,\hat{\theta}(t)) \\ \hat{x}^T(t/t)H^T(t,\hat{\theta}(t)) \end{bmatrix} P(t/t)dt \qquad (18b)$$

with initial conditions $\hat{x}(t_o/t_o) = E[x(t_o)]$, $\hat{\beta}(t_o) = p_1$, $\hat{\theta}(t_o) = E[\theta]$ and $P(t_o/t_o) = P(t_o)$; and where $P(t/t)$ is the error covariance matrix of the estimate, namely

$$P(t/t) = E\left\{\begin{bmatrix} x(t) - \hat{x}(t/t) \\ \theta - \hat{\theta}(t) \\ \beta - \hat{\beta}(t) \end{bmatrix} \begin{bmatrix} x(t) - \hat{x}(t/t) \\ \theta - \hat{\theta}(t) \\ \beta - \hat{\beta}(t) \end{bmatrix}^T /t\right\}$$

Proof: The proof consists of application of the relinearized equations to the linear model of section III, after augmenting the state vector as in (6a) and (6b), and therefore will be omitted.

In the following, two special cases of the corollary are given solely for illustrative purposes.

Case I: Estimation/System Identification

Let $\beta = 1$, θ, a scalar parameter and the model equations have the following form:

$$\begin{bmatrix} \dot{x}(t) \\ \theta \end{bmatrix} = \begin{bmatrix} \theta x(t) \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t)$$

$$z(t) = x(t) + v(t)$$

Then the relinearized Kalman-Bucy filter equations for the optimal estimates, $\hat{x}(t/t)$ and $\hat{\beta}(t)$ and estimation error covariance matrix are given by the following

$$\begin{bmatrix} dx(t/t) \\ d\hat{\theta}(t) \end{bmatrix} = \begin{bmatrix} \hat{\theta}(t)\hat{x}(t/t) \\ 0 \end{bmatrix} dt + P(t/t) \begin{bmatrix} I \\ 0 \end{bmatrix} R^{-1}(t)[dz(t) - \hat{x}(t/t)dt]$$

$$P(t/t) = \begin{bmatrix} \hat{\theta}(t)I & \hat{x}(t/t) \\ 0 & 0 \end{bmatrix} P(t/t)dt + P(t/t)\begin{bmatrix} \hat{\theta}(t)I & 0 \\ \hat{x}^T(t/t) & 0 \end{bmatrix} dt + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} dt -$$

$$P(t/t)\begin{bmatrix} R^{-1} & 0 \\ 0 & 0 \end{bmatrix} P(t/t)\,dt$$

As one can easily see upon examining the above equations, the highest conditional moments of $x(t)$ and θ that are required to obtain $x(t/t)$ and $\theta(t)$ are second order moments. Instead, the differential equation for optimal estimates require that the lower order moments equations contain terms which are functions of higher order moments, resulting in infinite dimensional problem.

Case II: Joint Detection-estimation

Let $\theta = 1$ and β, unknown, and let this joint detection-estimation problem take the following form

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\beta} \end{bmatrix} = \begin{bmatrix} x(t) \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t)$$

$$z(t) = \beta x(t) + v(t)$$

where $\beta = 1$ or 0. Then the relinearized Kalman-Bucy differential equations for the optimal estimates and error covariance matrix are given by the following equations.

$$\begin{bmatrix} d\hat{x}(t/t) \\ d\beta(t) \end{bmatrix} = \begin{bmatrix} \hat{x}(t/t) \\ 0 \end{bmatrix} dt + P(t/t)\begin{bmatrix} \hat{\beta}(t)\ I \\ \hat{x}^T(t/t) \end{bmatrix} R^{-1}(t)[dz(t) - \hat{\beta}(t)\hat{x}(t/t)\,dt]$$

$$dP(t/t) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} P(t/t)dt + P(t/t)\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} dt + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} dt - P(t/t)\begin{bmatrix} \hat{\beta}(t)\ I \\ \hat{\beta}(t)\hat{x}^T(t/t) \end{bmatrix}$$

$$R^{-1}\begin{bmatrix} \hat{\beta}(t)\ I \\ \hat{\beta}(t)\hat{x}(t/t) \end{bmatrix} P(t/t)\,dt$$

The comments given previously in Case I also apply here.

V. CONCLUSIONS

Recent results of Middleton and Esposito [1968] and Lainiotis [1969] single-shot joint detection-estimation for discrete data are extended to the single-shot continuous data case and generalized to joint Bayesian detection-estimation- and system identification, as well as to causal estimators. Specifically, it is shown that the above problem constitutes a class of nonlinear mse estimation problems, with the attendant difficulties in realizing the optimal nonlinear estimators. However, by utilizing the adaptive approach, namely by considering (θ, β) as an unknown parameter to be learned, closed form expressions are given. These are given in terms of the generalized likelihood ratio $\Lambda(t,\theta)$ each of which is the LR for testing for the model specified the parameter value θ. In essence then, is has been shown that system identification is equivalent to multihypothesis testing, with a continuum or finite sequence of hypotheses, respectively for continuous or finite discrete range of θ.

In addition to the integral expressions for $\hat{\theta}(t)$, $\hat{s}(t/t)$ and $\hat{\beta}(t)$, stochastic differential equations have been obtained describing the temporal evolution of $\hat{\beta}(t)$, $p_1(\theta/t)$ and $p(\theta/t)$. Moreover for the special case of linear models stochastic differential equations are obtained for the temporal evolution of $\hat{\beta}(t)$, $p_1(\theta/t)$, $\hat{x}_1(t/t, \theta)$, $P_1(t/t, \theta)$ and $p(\theta/t)$. Finally, s simple, suboptimal nonlinear estimation algorithm is given, based on the relinearized Kalman filter.

REFERENCES

Duncan, E.E., "Probability Densities for Diffusion Processes", Technical Report 7001-4, Stanford University, 1967

Esposito, R., [1969], Private Communication

Frost, P., [1970], "The Innovation Process and Its Application to Nonlinear Estimation and Detection of Signals in Additive White Noise," Proceedings of the UMR-M.J. Kelly Communications Conference, Rolla, Missouri, October 1970

Jazwinski, A.H. Stochastic Processes and Filtering Theory, Academic Press, New York, 1970

Kailath, T., [1969], "A General Likelihood-ratio Formula for Random Signals in Gaussian Noise", IEEE Transactions on Information Theory, IT-15, pp 350-361

Lainiotis, D.G., [1969], "On a General Relationship Between Estimation, Detection and the Bhattacharya Coefficient", IEEE Transactions on Information Theory, IT-15, pp 504-505

Licht, B.W., "Approximations in Optimal Nonlinear Filtering", Ph.D. Thesis, Case-Western Reserve University, January 1970

Middleton, D. and Esposito, E. [1968], "Simultaneous Optimum Detection and Estimation of Signals in Noise", IEEE Transactions on Information Theory, IT-14, pp 434-444

Lainiotis, D.G., [1971], "Joint Detection, Estimation and System Identification", Information and Control Journal, August 1971

MATHEMATICAL MODELS FOR WARM AIR HEATING SYSTEMS

ADEL ELTIMSAHY
University of Toledo
Toledo, Ohio 43606

LOUIS F. KAZDA
University of Michigan
Ann Arbor, Michigan 48104

ABSTRACT

In previous papers (5),(7) the authors focused their attention on applying optimal control techniques to a class of warm-air heating systems. They assumed that the mathematical models were already established. This paper is directed toward presenting in some detail the work that resulted in the models which were ultimately used in the earlier papers.

INTRODUCTION

Since an optimal heating system is expected to control human comfort, a comprehensive investigation was undertaken to establish those factors which contribute to human comfort in domestic heating systems. Results of this investigation revealed that the following items are major contributors to human comfort:

1. Air Temperature: The most important and most commonly recognized factor which contributes to comfort. An acceptable range is 70 to 75 degrees F.
2. Temperature Uniformity and/or Temperature Gradients: Almost as important as the value of the temperature itself. Spatial variation should be as small as possible for maximum comfort.
3. Air Motion: An important factor in reducing partial temperature variations. Although small values of air velocity will be tolerated, large values cannot be due to general discomfort and excessive evaporative heat loss. Acceptable values are 10 to 45 fpm.
4. Humidity: An important factor for both health and safety. Low values are more generally tolerated than high values. Acceptable values are 20 to 60 per cent.
5. Radiation: An important fundamental aspect of some heating systems. For gas-fired warm-air systems, it is perhaps desirable to increase the control temperature to counteract the increased body heat loss which occurs by radiative cooling of the human body to the inside surfaces of external walls when outdoor air temperature decreases.

Studies revealed that the functional relation between the above quantities varied from one individual to the next, and that no precise criterion could be found which could be used as a cost functional in the heating system optimization studies. It was decided, therefore, that in this study the criterion to be utilized for judging system improvements would be based on room temperature. That is, an improved system is one in which the room temperature deviation from some desired nominal value is smaller (in absolute value) over a given period of time. The reasons for this choice lie in its comparative analytical simplicity and the fact that room temperature has the greatest influence on human comfort.

COMPONENTS OF THE DOMESTIC HEATING SYSTEM

In order to study the control problem associated with a forced-air gas-fired domestic heating system it is necessary to investigate the various elements which constitute present-day heating systems. They can be summarized as follows:

1. Domestic Space - the region whose comfort is to be controlled.
2. Room Boundaries - the walls, ceiling, and floors.
3. Heat-exchanger - that part of the furnace which supplies thermal energy to the domestic space.

4. Room Thermostat - the logic unit which supplies signals to the furnace for control.
5. Hot-air ducts - to transfer the heated air from the furnace to the domestic space.
6. Cold-air ducts - to transfer air from the domestic space to the furnace.
7. The blower and gas valve as parts of the heat exchanger control.

Since comfort can only be controlled when the blower is running, the analysis of heat exchanger assumes that a given amount of air/min. is being forced through it. In addition the time-constant of the gas-control valve was found to be negligible compared to the thermal time-constants of the system and was therefore neglected. A block diagram of system is shown in Figure 1.

THE ROOM MODEL

The room model as discussed herein refers to the study of the physics of a domestic room space under the conditions of thermal forced air convection in which the room space is assumed to possess a non uniform temperature distribution. Realizing that the room is a complex distributed parameter system, the object of this study was to develop a lumped parameter mathematical model which possesses the predominate features of the distributed parameter system but which can be used as an element of an optimal heating system study. The room is divided into two main parts: a) the domestic space, b) the walls, ceiling and floor. From a knowledge of the physics of the various elements that compose the room system, the mathematical model of each element is determined and then combined together to form the total model--one that would adequately define the dynamic behavior in the operating range of interest.

Let us consider a room with height H, width W, and length L. The room has three inside walls and one outside wall. An outside wall is one which possesses an exterior exposure to the climatic elements. Let the rest of the house be at a temperature range which is equal to the temperature range of this room. This means that there will be no heat transfer through the three inside walls or the floor or the ceiling. Heat is transferred to the outside only through the outside wall. Let the air inlet to the room be a rectangular opening and let the jet discharge parallel to the inside wall with one edge of the outlet coinciding with the inside wall. See Figure 2.

Heat transfer by radiation will be neglected with respect to the quantities transferred by conduction and convection. This is justified since no radiating objects are assumed in the room and the quantity of heat radiated from the sun will be included as an outside source.

Since the jet is adjacent to the wall when it expands into the room, its form will be a semi cone shape. The temperature of an element volume of air coming out from the register will decrease gradually according to a certain variation mainly due to the cooler air already present in the room. At a certain height (a character of the jet and the height of the room will be specified later), the semi-conical form of the jet disappears and air circulation in the room starts. Based on some previous experimental work on air distribution in the room,(1) the following regions (shown in Figure 2) will be used to establish the model. The choice of the three regions as seen is based on the way the air distributes itself in the room.

Region 1 or the Jet Region Model

In an actual room the airstream is not free since it is influenced by walls, ceilings, floors, and obstruction. However, tests have shown that(2) in cases where the outlet area itself is small compared to the dimensions of the space normal to the jet, the jet may be considered free as long as:

$$x \leq 1.5 \sqrt{A_s} \qquad (1)$$

where x = Distance from face of outlet in feet

A = Cross sectional area of the confined space in square feet.

We shall assume that this inequality is true for any x in the jet region; because of its location adjacent to a wall, the free jet analysis applies without any appreciable error. In general the velocity profile at any cross-section of a free jet can be approximated by a normal probability curve. However, due to the existence of the wall beside the air register, the velocity profile for the stream will not be exactly half a normal probability curve. The velocity of the air at the surface of the wall will be practically zero; but for engineering purposes, the normal probability curve can be used in order to calculate average temperatures over the jet cross section without an appreciable error. Experimentally it was found that:(2)

$$E_x = \frac{1.4x}{3\sqrt{A_o}} \qquad (2)$$

where E_x = entrainment ratio, $\frac{A_o}{2}$ = area of inlet in ft^2

Define now the following:

$T_i(t)$ = average temperature of air when leaving the jet to region 2
$T_3(t)$ = average temperature of air in region 3 and the room initially
Q_o = volume flow of air in ft^3/min. at inlet of the jet region
Q_x = volume flow of air in ft^3/min. at a distance x
Q_1 = volume flow of air in ft^3/min. when leaving the jet region to region 2
ρ = density of air in lbs./ft^3
C_p = specific heat of air at constant pressure in BTU/(lb)(F).
τ_x = time taken by a unit volume of air to travel from the inlet for a distance x in minutes
τ_1 = time taken by a unit volume of air to travel through the whole jet region in minutes

As a first step here, the average temperature at a distance x will be calculated as a function of x. If it is assumed that the air is coming in the room with a temperature T_i, the average temperature at a distance x is T_x, and that there are no losses through the inside wall, then a heat balance equation can be written for the corresponding control volume to distance x. The control volume in thermodynamics is defined as a volume in space through which matter flows, and it is bounded by a control surface. Consider now the control volume corresponding to a distance x in the jet region. From the application of the first law of thermodynamics to a thermodynamic open system knowing that there is no work done in the process and neglecting the kinetic and potential energies associated with the moving air, the following heating balance equation can be written:

$$Q_x T_x(t) = Q_o T_i(t-\tau_x) + (Q_x - Q_o) T_3(t) \qquad (3)$$

or

$$T_x(t) = \frac{3\sqrt{A_o}}{1.4}\,[T_i(t-\tau_x) - T_3(t)]\,\frac{1}{x} + T_3(t) \qquad (4)$$

Now taking the whole jet region as one control volume we have:

$$T_1(t) = \frac{3\sqrt{A_o}}{1.4}\,[T_i(t-\tau_1) - T_3(t)]\,\frac{1}{x} + T_3(t) \qquad (5)$$

Due to the small range of the expected variations of the absolute temperature, the air density and its specific heat are taken to be constants all over the space of the room. Since the time delay in the jet region is very small compared with the rest of the system, the equation of this region takes the following form:

$$Q_1T_1(t) = Q_oT_i(t) + (Q_1-Q_o)T_3(t) \tag{6}$$

Region 2 Model

The whole volume of region 2 will be assumed to be at a state of complete mixing.(1) This allows a uniform temperature assumption in this region. Call this temperature T_2. From the study of air movement in jets,(2) the effect of the ceiling will not be significant until the jet approaches the ceiling to within half a jet diameter. At height H, the diameter of the jet will be 0.4H. Therefore,

$$X = \text{length of the jet region in ft.} = 0.8\ H$$

Applying now the first law of thermodynamics to the control volume of region 2, and taking into consideration that there will be a rate of change of internal energy in this case due to the heat transfer to the outside wall we get:

$$Q_1\rho C_pT_1(t-\tau_2) - Q_1\rho C_pT_2(t) - k_2[T_2(t)-T_{w_2}(t)] = \rho C_pV_2\frac{dT_2(t)}{dt} \tag{7}$$

where:

$T_{w_2}(t)$ = average temperature of the inside surface of the part of the outside wall facing region 2 in F°

τ_2 = time taken by a unit volume of air to travel through region 2 in minutes

k_2 = average proportionality constant (or surface coefficient) defining the heat transfer by convection from the air in region 2 to the inside surface of the outside wall facing region 2

V_2 = volume of air in region 2 = (wL)(0.2H) = 0.2 wLH ft^3

Again it was found that the time delay τ_2 of region two can be neglected with respect to the rest of the system. Therefore:

$$\dot{T}_2(t) = \alpha_1T_1(t) - \beta_1T_2(t) + \gamma_1T_{w_2}(t) \tag{8}$$

where

$$\alpha_1 = \frac{Q_1}{V_2},\ \beta_1 = \frac{k_2+Q_1\rho C_p}{\rho C_pV_2},\ \gamma_1 = \frac{k_2}{C_pV_2}$$

Region 3 Model

Again complete mixing is expected in this region.(1) The heat input to this region will be from region 2 by convection, and to the cold air duct. Applying now the first law of thermodynamics to region 3 as we did for region for region 2 we get:

$$Q_1\rho C_pT_2(t-\tau_3)-Q_1\rho C_pT_3(t)-k_3[T_3(t) - T_{w_3}(t)] = \rho C_pV_3\ \frac{dT_3(t)}{dt} \tag{9}$$

where

V_3 = volume of air in region 3 in ft^3

τ_3 = time taken by a unit volume of air to travel through region 3 in minutes

T_{w_3} = average temperature of the inside surface of the outside wall facing the third region in F°

k_3 = average proportionality constant defining the heat transfer by convection to the inside surface of the outside wall facing the third region

By using Taylor's series expansion of $T_2(t-\tau_3)$, and taking only the first order approximation, therefore, the state variable Equation (9) of the third region takes the following form:

$$\dot{T}_3(t) = \alpha_2 T_1(t)+\beta_2 T_2(t)-\gamma_2 T_3(t)-\delta_2 T_{w_2}(t)+\zeta_2 T_{w_3}(t) \qquad (10)$$

where α_2, β_2, γ_2, δ_2 are ζ_2 are parameters of region 3.

THE WALL MODEL

In constructing the wall model, (see Figure 3a and 3b) thermal circuit techniques were used. Assuming that heat flows only in one direction through the wall, the wall behaves as a distributed parameter RC transmission line. As in transmission lines, it can be shown that the errors in representing the wall under the same boundary conditions varies inversely with the square of the number of T circuits used.(3)

A compromise between the degree of approximation and overall cost of the optimum controller is made, since this puts a preliminary bound on the cost of the optimum controller. The derivation for the wall model follows: it is divided into two parts; (a) facing region 2 (b) facing region 3.

Consider now Figure 3b which represents the thermal circuit of a one T network wall. The outside end of the wall which is exposed to the sun and wind is equivalent to a known temperature source T_o (the effective air temperature). R_o represents the resistance whose conductance represents the heat flow between the outside surface of the wall and the surrounding atmosphere. The inside of the wall is facing the room heat capacity. R_i is the resistance whose conductance represents heat transfer between the air in the domestic space and the inside surface of the wall. R and C are the equivalent resistance to heat transfer and heat capacity of the wall respectively. They are functions of the properties of the material of the wall.

Let V be the temperature of the air inside the room; V_1 be the temperature of the inside surface of the wall; V_2 be the temperature of the outside surface of the wall. It follows that:

$$-CR(R+R_o)\frac{dV}{dt} + C(R_i+R)(R+R_o)\frac{dV_1}{dt} - (2R+R_o)V+(2R+R_o+R_i)V_i-R_iE_o=0 \qquad (11)$$

Using Equation (11), the state variable equations for parts of the wall facing regions two and three are stated in the following, where subscripts 2 and 3 refer to regions two and three respectively.

For Region 2:

In this case T_2 corresponds to V and T_{w_2} corresponds to V_1. Therefore:

$$-C_2R_2(R_2+R_{o2})\frac{dT_2}{dt} + C_2(R_{i_2}+R_2)(R_2+R_{o2})\frac{dT_{w_2}}{dt} - (2R_2+R_{o2})T_2 + (2R_2+R_{o2}+R_{i2})T_{w_2} - R_{i2}T_o = 0 \qquad (12)$$

For Region 3:

Again T_3 corresponds to V and T_{w_3} corresponds to V_1:

$$-C_3R_3(R_3+R_{o3})\frac{dT_3}{dt} + C_3(R_{i_3}+R_3)(R_3+R_{o3})\frac{dT_{w_3}}{dt} - (2R_3+R_{o3})T_3 + (2R_3+R_{o3}+R_{i3})T_{w_3} - R_{i3}T_o = 0 \qquad (13)$$

Substituting Equation (8) into Equation (12), transposing and rearranging gives:

$$\dot{T}_{w_2} = -\alpha_3 T_{w_2} - \beta_3 T_1 - \gamma_3 T_2 + \delta_3 T_o \tag{14}$$

Similarly using Equations (10) and (13), gives:

$$\dot{T}_{w_3} = -\alpha_4 T_{w_3} + \beta_4 T_1 - \gamma_4 T_2 - \delta_4 T_3 - \zeta_4 T_{w_2} + \eta_4 T_o \tag{15}$$

where α_3, γ_3, δ_3, α_4, β_4, γ_4, δ_4, ζ_4, and η_4 are fixed parameters of the system.

THE ROOM SYSTEM MODEL

Summarizing, the set of equations representing the room space and walls is:

$$T_1 = \frac{Q_o}{Q_1} T_i + (1 - \frac{Q_o}{Q_1})T_3$$

$$\dot{T}_2 = \alpha_1 T_1 - \beta_1 T_2 + \gamma_1 T_{w_2}$$

$$\dot{T}_3 = -\alpha_2 T_1 + \beta_2 T_2 - \gamma_2 T_3 - \delta_2 T_{w_2} + \zeta_2 T_{w_3}$$

$$\dot{T}_{w_2} = -\alpha_3 T_{w_2} - \beta_3 T_1 - \gamma_3 T_2 + \delta_3 T_o$$

$$\dot{T}_{w_3} = -\alpha_4 T_{w_3} + \beta_4 T_1 - \gamma_4 T_2 - \delta_4 T_3 - \zeta_4 T_{w_2} + \eta_4 T_o$$

This is a set of linear first order differential equations with constant coefficients. This set represents the room system, T_i is the input, T_1, T_2 or T_3 may be considered as the output of the system, and T_o the disturbance. In matrix form these equations may be written as:

$$\begin{bmatrix} \dot{T}_2 \\ \dot{T}_3 \\ \dot{T}_{w_2} \\ \dot{T}_{w_3} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} T_2 \\ T_3 \\ T_{w_2} \\ T_{w_3} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} T_i + \begin{bmatrix} 0 \\ 0 \\ m_1 \\ m_2 \end{bmatrix} T_o \tag{16}$$

CONCLUSIONS

The room model developed in this paper was used as a basis to study the characteristics of domestic forced air heating systems. A comparison of the computer simulation results with experimental studies is shown in Figures 4, 5, and 6. As a result of this study it was concluded that for most domestic heating systems, region 3 constituted the major portion of the habitable room. Considerable simplification in the number of state-variables needed to characterize the system could be achieved, if the room space were represented by a single temperature, T_R. Letting $T_1 = T_2 = T_3 = T_R$ and $T_{w_2} = T_{w_3} = T_w$, the approximate mathematical model of the room now assumes the following form

$$\rho C_p V \dot{T}_R = (\rho C_p Q\, T_i - \rho C_p Q\, T_R) - k(T_R - T_W) - CR(R+R_o)\dot{T}_R + C(R_i+R)(R+R_o)\dot{T}_W$$

$$-(2R+R_o)T_R + (2R+R_o+R_i)T_W - R_i T_o = 0$$

where

V = volume of the room
Q = rate of flow of air
k = average proportionality constant defining the heat transfer by convection to the inside wall
C = equivalent heat capacity of the outside wall
R = equivalent resistance of the outside wall
R_o = equivalent outside air to surface resistance of the outside wall
R_i = equivalent inside air to surface resistance of the outside wall

The above simplification is considered to be of considerable engineering significance, since it allows a reduction in the number of equations representing the parameters of the optimal law from 21 to 10. The above equations were used in the optimal studies presented in references 5 and 7.

REFERENCES

1. "Distribution of Air Within a Room for Year-Round Air-Conditioning," Parts I and II, Harold E. Straub, Stanley F. Gilman, Seichi Konozo, and Michael Ming Chen, University of Illinois Engineering Experiment Station Bulletin Nos. 435 and 442.

2. American Society of Heating and Refrigerating and Air-Conditioning Engineers, Guide and Data Book, 1963, Chapter 11.

3. The Study of the Thermal Behavior of Structures by Electric Analogy, G. Brunard, British Journal of Applied Physics, 1950.

4. "The Optimization of Domestic Heating Systems," Adel H. Eltimsahy, Ph.D. Dissertation, April 1967, University of Michigan, Ann Arbor, Michigan.

5. A Sub-Optimal Heating System, Adel H. Eltimsahy and Louis F. Kazda. Proc. of Allerton Conference, October 1968.

6. "The Warmed-Air Domestic Space Heating System - Considered As a Closed Loop Control System," Michael Burrow and Louis F. Kazda, ORA Final Report Project 03506, University of Michigan.

7. "An Optimal Gas-Fired Heating System," A. H. Eltimsahy and L. F. Kazda, "Automatica" Vol. 5, pp. 741-754. Pergamon Press, 1969; also Proc. of 4th IFAC Congress, Warsaw, Poland, 1969.

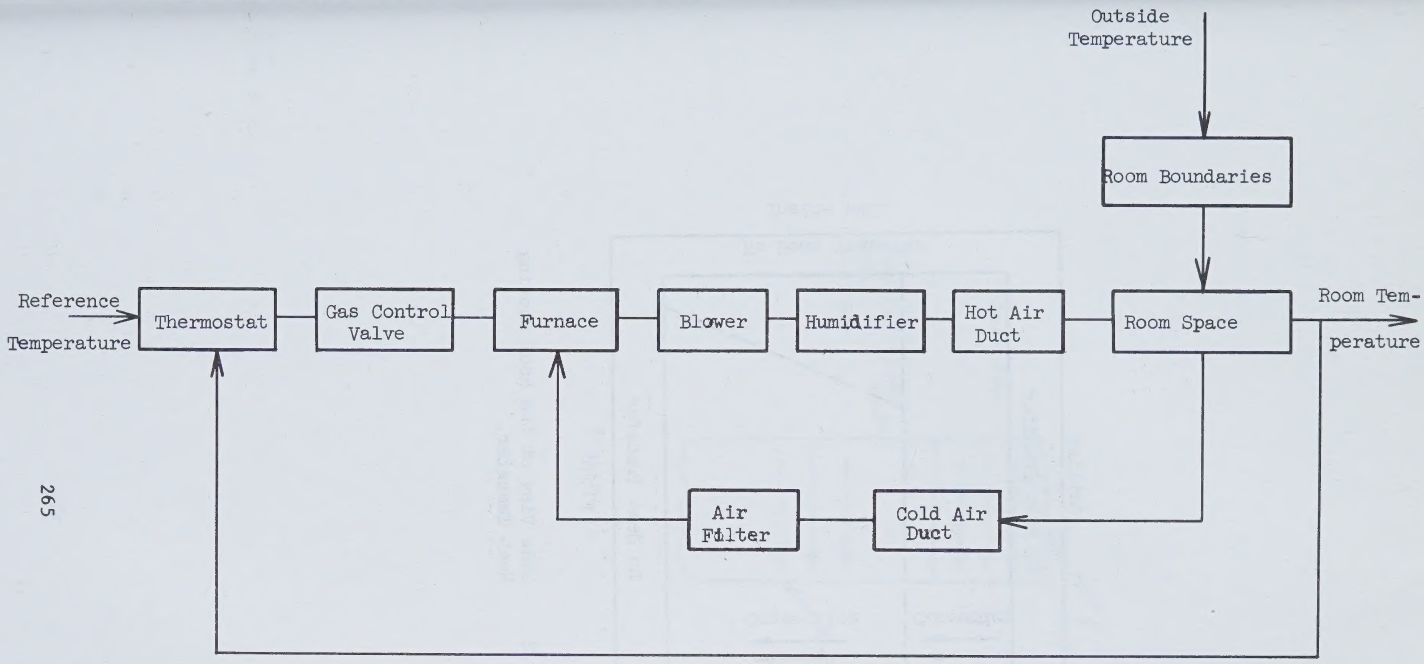

Outside
Temperature
Room Boundaries
Reference
Temperature
Thermostat
Gas Control
Valve
Furnace
Blower
Humidifier
Hot Air
Duct
Room Space
Room Tem-
perature
Air
Filter
Cold Air
Duct

Figure 1 Components of the Heating System.

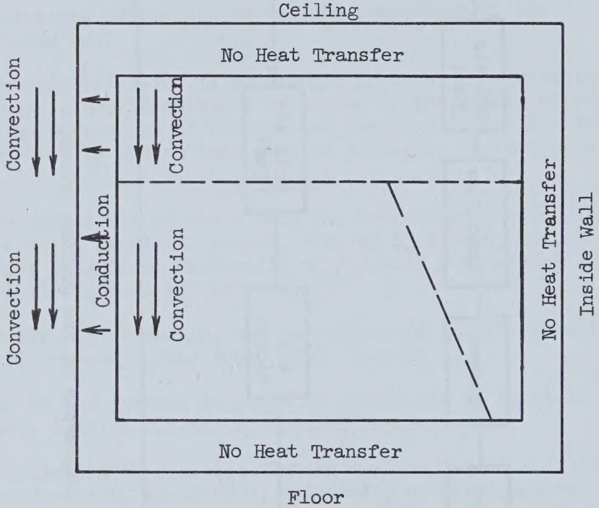

Ceiling
No Heat Transfer
Convection
Convection
Conduction
Convection
Convection
No Heat Transfer
Inside Wall
No Heat Transfer
Floor

Figure 2 Side View of the Room Showing Heat Transfer.

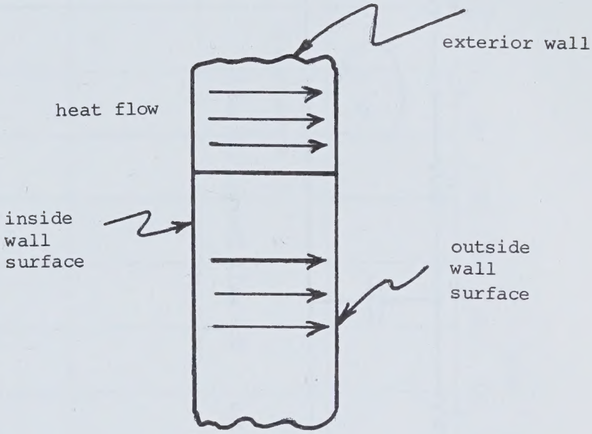

exterior wall
heat flow
inside wall surface
outside wall surface

Figure 3.a

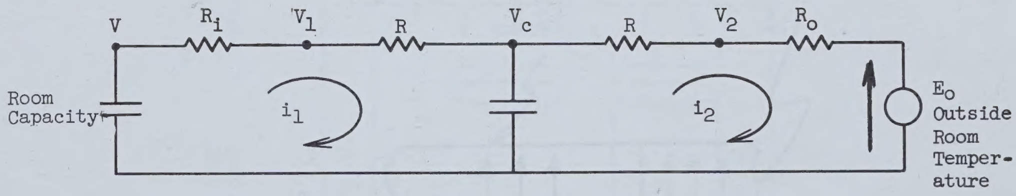

V
R_i
V_1
R
V_c
R
V_2
R_o
Room
Capacity
i_1
i_2
E_o
Outside
Room
Temper-
ature

Figure 3.b The Wall Thermal Circuit.

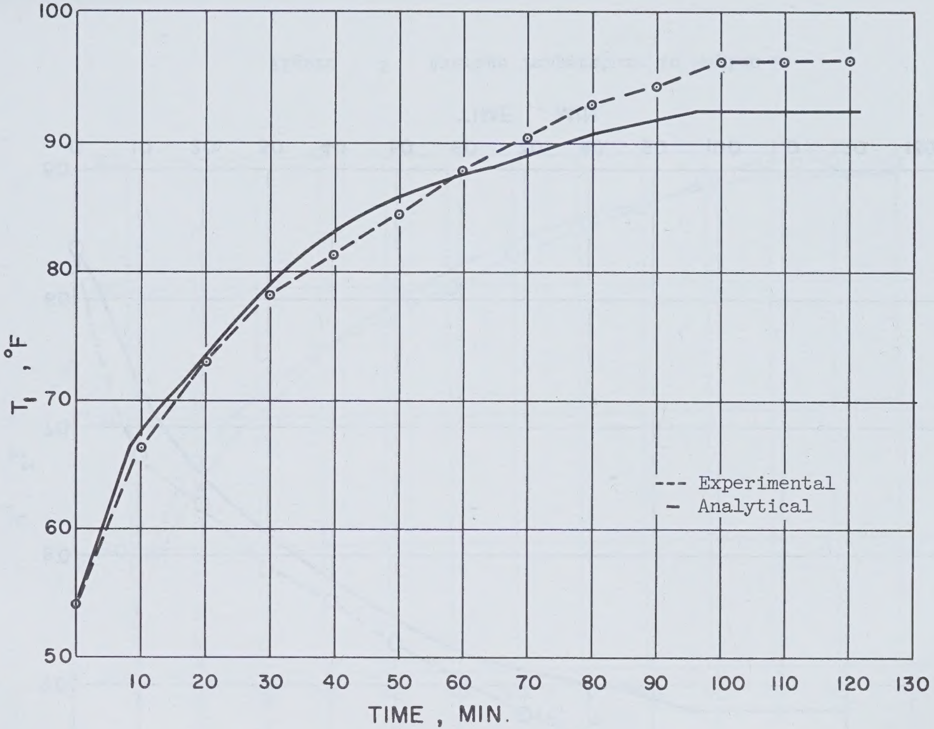


Figure 4 Average Temperature in Region 1.

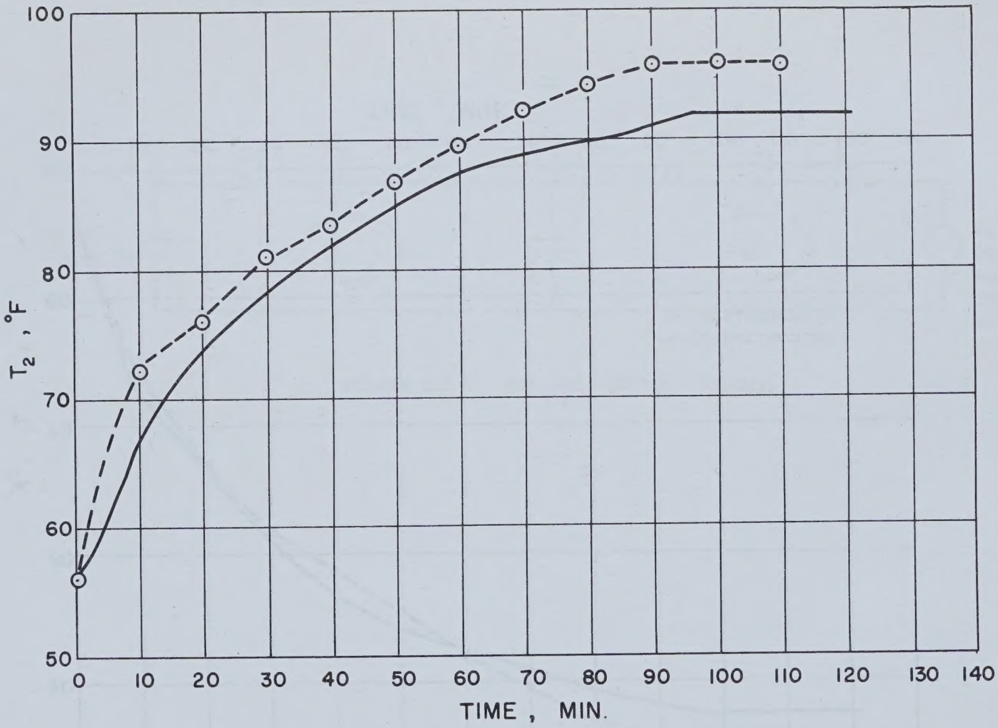


Figure 5 Average Temperature in Region 2.

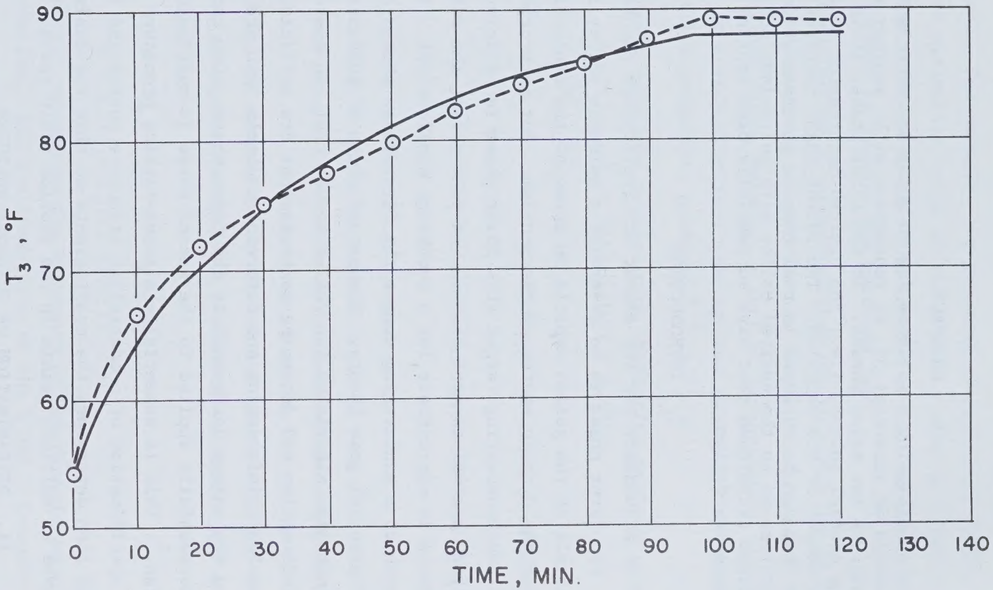


Figure 6 Average Temperature in Region 3.

APPLICATION OF MODERN CONTROL THEORY TO GUIDANCE LAW DESIGN

DAVID M. SALMON
Systems Control, Inc.
Palo Alto, California

ABSTRACT

The process of designing a guidance law to guide one vehicle (the interceptor) towards an intercept of, or rendezvous with, another vehicle (the target) involves two major phases. In the first phase, it is necessary to determine the control policy, i.e., the desired interceptor control as a function of the states of the vehicles. The second phase requires synthesis of an algorithm that can be employed in real time to generate guidance commands that correspond to the desired policy with sufficient accuracy. The paper describes techniques that have successfully been applied to both phases of guidance law design.

I. INTRODUCTION

The design of a guidance law for almost any application can be considered in two phases. The first phase is to determine a guidance policy i.e., to specify the controls of the guided vehicle in terms of the available information about the dynamic states of the problem. For intercept or rendezvous of a non-maneuvering target this first phase is efficiently treated by applying optimal control theory. In particular, the method of predictive guidance is appropriate for a number of applications. For intercept or rendezvous of a maneuvering target the first phase should be treated by applying differential game theory. The second phase of guidance law design is to synthesize algorithms for real-time use that can operate on the available information and determine controls that are sufficiently close to the available information and determine controls that are sufficiently close to the policy determined in the first phase. One technique that has been successfully applied to the second phase is multipoint controller design.[1] This is essentially a curve-fitting procedure that first requires specification of a controller structure containing free coefficients and then determines the coefficients so that the desired policy is achieved at multiple points in the problem state space.

II. DETERMINATION OF GUIDANCE POLICIES

A. Predictive Guidance for Intercept or Rendezvous of a Non-Maneuvering Target

The intercept or rendezvous of a non-maneuvering target vehicle (T) with an intercept vehicle (I) can be formulated as a nonlinear stochastic optimal control problem. Standard practice in solution of this problem is to enforce a separation of estimation and control i.e., (1) generate estimates $\hat{x}_T(t|t)$ and $\hat{x}_I(t|t)$ of the T and I states at time t given measurements to time t, and (2) employ these estimates in a guidance law designed using a deterministic analysis. In many cases this approach can be justified after the fact. For example, the greatest contribution to the terminal cost may be due to estimation error of the T state, with no possibility of improving this estimate by altering the I control. In such a case it is of no practical value to improve the guidance law.

Even if estimation is separated from the guidance law design the remaining nonlinear deterministic optimization problem is generally formidable. In many practical cases, however, a type of linearization can be performed which leads to a very effective near-optimal policy. The resulting guidance policy is called predictive guidance.[2,3]

The dynamic equations for deriving predictive guidance are:

$$\dot{x}_I = v_I, \quad \dot{v}_I = T(t) - D(t) + a_I(t), \qquad \dot{x}_T = v_T, \quad \dot{v}_T = a_T(t), \tag{1}$$

where x_I, x_T are the 3 dimensional Cartesian position vectors for I and T,

v_I, v_T are the 3 dimensional Cartesian velocity vectors for I and T

$T(t)$ is the thrust applied to I

$D(t)$ is the drag applied to I

$a_I(t)$, $a_T(t)$ are the 3 dimensional Cartesian acceleration vectors for I and T.

The vector $a_I(t)$ is assumed to be the control for I. In the following it is convenient to employ the relative position, velocity, and acceleration vectors defined by

$$x = x_T - x_I, \quad v = v_T - v_I, \qquad a = a_T(t) - [T(t) - D(t)] \quad . \tag{2}$$

Equation (1) recognizes the structure of the dynamic equations for a guidance problem. All nonlinearities of the T and I dynamic equations have been expressed as functions of time and incorporated into the acceleration vectors T(t), D(t), and a(t). A convenient form of performance index is:

$$J = \frac{1}{2} c_1 x'(t_f)\, x(t_f) + \frac{1}{2} c_2 v'(t_f)\, v(t_f) + \frac{1}{2} \int_{t_o}^{t_f} a_I'(t)\, a_I(t) dt \tag{3}$$

where t_f is the terminal time (usually free), and c_1, c_2, are scalar weighting factors.

Equations (1) and (2) are in the form of a standard optimal control problem which can be solved in closed form. After a considerable amount of algebra, the solution can be written as follows.[4]

$$a_I(t) = g_1(t)\ x(t_f,t) + g_2(t)\ v(t_f,t) \tag{4}$$

where the scalars $g_1(t)$ and $g_2(t)$ are

$$g_1(t) = \frac{\frac{1}{c_2}(t_f-t) + \frac{1}{2}(t_f-t)^2}{\frac{1}{c_1c_2} + \frac{1}{c_1}(t_f-t) + \frac{1}{3c_2}(t_f-t)^3 + \frac{1}{12}(t_f-t)^4}$$

$$g_2(t) = \frac{\frac{1}{c_1} - \frac{1}{6}(t_f-t)^3}{\frac{1}{c_1c_2} + \frac{1}{c_1}(t_f-t) + \frac{1}{3c_2}(t_f-t)^3 + \frac{1}{12}(t_f-t)^4} \tag{5}$$

and the 3-vectors $x(t_f,t)$ and $v(t_f,t)$ are

$$x(t_f,t) = x(t) + v(t)(t_f-t) + \int_t^{t_f}\int_t^{\sigma} a(\tau)d\tau\ d\sigma$$

$$v(t_f,t) = v(t) + \int_t^{t_f} a(\tau)d\tau \quad . \tag{6}$$

The vectors $x(t_f,t)$ and $v(t_f,t)$ are readily identified as the predicted miss-distance and miss-velocity of T and I at time t_f.

These results are intuitively appealing. The optimal control at time t is to accelerate I in a direction which is a weighted average of the predicted miss-distance and miss-velocity at time t_f. For rendezvous, with a non-negligible weight on the integral control cost the correct time-varying constants of proportionality are given by Eq. (5). If no weight is desired on the control cost, c_1 and c_2 may be allowed to increase without limit, yielding

$$a_I(t) = \frac{6}{(t_f-t)^2}\ x(t_f,t) - \frac{2}{(t_f-t)}\ v(t_f,t) \quad . \tag{7}$$

For intercept problems with no cost on terminal velocity-miss the constant c_2 is set equal to zero, yielding

$$a_I(t) = \frac{c_1(t_f-t)}{1 + \frac{c_1}{3}(t_f-t)^3}\ x(t_f,t) \quad . \tag{8}$$

If in this case no weight is desired on the integral cost term then c_1 may be allowed to increase without limit yielding

$$a_I(t) = \frac{3}{(t_f-t)^2}\ x(t_f,t) \quad . \tag{9}$$

As an example application of predictive guidance consider the interception of a non-maneuvering reentry vehicle by an interceptor having lateral control achieved aerodynamically (i.e., by use of control fins to alter the vehicle attitude). A typical procedure is as follows.

1. The T trajectory $\hat{x}_T(t|\sigma)$ at future times, t, is predicted based on information to the present time, σ.
2. The I trajectory $\hat{x}_I(t|\sigma)$ is also predicted forward, using the optimal control computed at the previous guidance update, i.e., $a_I(t)$ as obtained from Eq. (8) using the t_f and $x(t_f,t)$ obtained at the previous update.
3. A search is made over the terminal times t_f to find that terminal time $\hat{t}_f$ yielding a predicted miss-distance $x(\hat{t}_f,\sigma)$ which allows Eq. (8) to be satisfied (while taking into account the achievable control accelerations for I).
4. The optimal control is obtained from Eq. (8), using $\hat{t}_f$ and $x(\hat{t}_f,\sigma)$ in place of t_f and $x(t_f,t)$. This control is applied until the guidance is again updated.

Such a procedure can, of course, be used as the basis of a real-time guidance law. However, the requirement for continually predicting the T and I trajectories (usually by integrating the nonlinear dynamic equations) leads to a cumbersome law requiring a considerable amount of real-time computation. Accordingly it is considered here that predictive guidance merely provides a desirable control policy, i.e. it defines a desired control for each possible state of T and I.

B. Determination of Guidance Policies for a Goal-Tending Game

A brief description of the formulation and solution of a pursuit-evasion problem which is representative of a class of goal-tending games is given in this section. Details are given in Ref. 5. A slow Pursuer, P, and a faster Evader, E, move in three-dimensional space. P desires to protect a defended site from the attacker E, who desires to approach the site as closely as possible without being captured by P. Capture of E occurs of the distance between P and E becomes less than a given distance, c. Figure 1 shows some of the important parameters of the game. The site to be defended by P is conveniently situated at the origin. P is initially at the origin and free to leave it at any time. E is initially a long way from the origin.

The state variables for E are the position coordinates x_e, y_e, z_e, and the angles ϕ and ψ shown in Fig. 1 that determine the present direction of motion. The speed of E is normalized to unity. The control variables for E are u, and θ, the magnitude and direction of the lateral acceleration (which can be changed instantly). The controls are limited by $0 \leq u \leq 1.0$, $-\pi \leq \theta \leq \pi$, implying that E, if turning in a plane can turn no sharper than on a circle of unit radius.

The state variables for P are the position coordinates x_p, y_p, z_p, the angles ξ_p and η_p shown in Fig. 1, and the variables v_p and δ_p which determine the magnitude and direction of the lateral acceleration. The speed of P is constant at a units, $a<1.0$. The control variables for P are two additional variables β_p and γ_p which determine respectively the rates of change of v_p and δ_p. The controls are constrained so that P, if turning in a plane, can turn no sharper than on a circle of given radius b, $b<1$, and so that the rate of change of the lateral acceleration is not greater than $\frac{1}{d}$. If, for example, P is initially travelling with zero acceleration there is a delay, d, before P can apply full lateral acceleration (given by $\frac{a^2}{b}$).

The game payoff to be minimized by E and maximized by P is the distance of closest approach from E to the origin.

The key to solving this differential game is the recognition that at some instant of time in an optimal play of the game, E must be on the verge of capture by P. If there is no such critical instant of time it follows that either the minimum distance from E to the origin is zero or that E could have reached a position closer to the origin without capture.

The conditions corresponding to E being on the verge of capture can readily be derived mathematically. These conditions imply that P is a distance c from E and that the velocity of P relative to E is tangential to the capture circle of radius c centered at the position of E. Following the critical time instant where E narrowly evades capture, P is no longer a threat to E since E travels faster, and the optimal control for E is simply to turn towards the origin as quickly as possible. As a result, the payoff (the minimum distance from E to the origin) can be expressed in terms of the position and direction of travel of E at the critical time. Thus, the game can be analyzed as a game ending at the critical time. Differential game theory[6,7] provides necessary conditions for the trajectories of P and E before the critical time, so that candidate optimal trajectories can be propagated backward in time starting from the verge-of-capture conditions. Candidate game solutions are then obtained by appropriately positioning these trajectories with respect to the origin. To verify that a local solution to the differential game has indeed been achieved, two one-sided checks can be performed. That is, using the determined strategy for E, determine if E can decrease the payoff; then using the determined strategy for E, determine if P can increase the payoff. Once the optimal strategies are known, the optimal game payoff (i.e., the size of site that P can guarantee to defend) can be computed from the geometry of the trajectories, and curves can be generated that show how the payoff varies with the game parameters (such as speed, maneuverability, and delay in response for P).

Figure 2 shows the nature of the solution when P can turn instantaneously (i.e. the state variables v_p, and δ_p are ignored, and η_p and ξ_p become control variables). E travels directly towards the origin and reaches E_{to} when P, having left the origin earlier, reaches P_{to}. If E delays maneuvering after this time then P will enter the capture region of E (i.e. P will eventually capture E if P responds correctly to E maneuvers). Accordingly E makes a maximum evasive maneuver (in any plane containing E_{to} and P_{to}). On detecting the E maneuver, P follows and narrowly misses capturing E when E is at $E_{\bar{t}}$ and P is at $P_{\bar{t}}$. Then E turns back towards the origin. The game payoff is the distance shown in Fig. 2.

The same qualitative form of solution has been found to apply for the more complex dynamics described above even when there are large delays, d, in the response of P. Of course, if P has a large delay in response, E only needs to perform a small "side-step" maneuver to pass safely by P, and therefore achieves a small payoff, D.

Based on these results and the obvious nature of the optimal strategies it is evident that the same methods can be applied to solve similar problems with more complex dynamics. It is probable that the same qualitative nature of solution also holds for many of these games.

These results of the above game only specify the optimal P and E trajectories. Thus the results only partially specify a policy for both P and E in that the desired controls for certain states are specified. However, it is relatively easy, once the optimal strategies are known, to determine satisfactory controls for P and E over the remainder of the state space (i.e. over regions corresponding to sub-optimality of the other player). See, for example, Ref. 8 which describes a technique for generating suitable P controls when P is inside the capture region of E.

III. SYNTHESIS OF A REAL-TIME GUIDANCE LAW

Once the guidance policy has been determined it is necessary to synthesize a guidance law that realizes the desired policy with adequate accuracy but requires little real-time computation. The multipoint controller design method has been applied to this synthesis. The method has four steps.

1. Representative trajectories are generated through the state space and the associated controls defined by the policy are determined.
2. A controller structure is postulated containing undetermined functions of states.
3. Data from the representative trajectories is used to determined the best values of the undetermined functions (a least squares fit is convenient).
4. The controller performance is checked over the state space of concern.

This technique has been applied to the problem of intercepting a ballistic (non-maneuvering) reentry vehicle with an aerodynamically controlled interceptor.[9] A satisfactory guidance policy for this problem is predictive guidance using Eq. 8. The following very simple "multipoint guidance" law has been found to provide adequate accuracy. The law is based on defining a "desired direction of travel", v_I^{des}, such that if the interceptor were travelling in that direction and applied zero future control it would intercept the target with zero miss-distance. An approximation of the predictive guidance policy can then be derived in terms of the angle $\Delta\phi$ between the desired and actual directions of travel. The magnitude of the control acceleration is simply

$$|a_I| = \frac{\bar{t}^2 |\hat{v}_I| \Delta\phi}{\frac{1}{c_1} + \frac{1}{3}\bar{t}^3} \qquad (12)$$

where c_1 is the weighting on miss-distance (Eq. 3)

$\bar{t}$ is an approximate value of time-to-go (which can be expressed in terms of the present T and I states)

$\hat{v}_I$ is the present estimate of interceptor velocity.

The acceleration is applied in the plane of the vectors $\hat{v}_I$ and v_I^{des} in order to correct $\hat{v}_I$ towards v_I^{des}. The desired direction of travel v_I^{des} lies in the plane defined by the present estimates of target velocity

$\hat{v}_T$ and interceptor position $\hat{x}_I$. The direction of v_I^{des} in this plane is given by the angle ϕ_{des}, defined in Fig. 3. It has been found that ϕ_{des} can be expressed in terms of the relative state variables of Fig. 3 as follows.

$$\phi_{des} = h_1(\hat{R})a + h_2(\hat{R})a|a| \tag{13}$$

where
$$a = \frac{v_T^{ave} \sin \hat{\delta}}{v_I^{ave}}$$

$$v_T^{ave} = \hat{v}_T - \frac{\hat{D}_T \overline{t}}{2}$$

$$v_I^{ave} = \hat{v}_I - \frac{\hat{D}_I \overline{t}}{2}$$

$\hat{D}_I, \hat{D}_T$ are respectively the current estimated drag forces for I and T.

The functions $h_1(\hat{R})$ and $h_2(\hat{R})$ are obtained in the design phase by performing a least squares fit to the predictive guidance policy along a number of trajectories which are representative of the geometries of intercept and types of reentry vehicle of concern.

Motivation for the particular guidance law structure can be seen by examining the form of ϕ_{des} when T and I have constant velocities, v_T and v_I. Namely,

$$\phi_{des} = \arc \sin \frac{v_T \sin\delta}{v_I} \quad . \tag{14}$$

Direct application of a law of the form of Eq. (14) does not yield adequate accuracy, but the law of Eq. (13) when tuned to the particular dynamics of T and I has been found to perform well (i.e. accurately emulate the predictive guidance policy) over the desired range of intercept geometries and T characteristics. The computation requirements of this law are less than a tenth of those for conventional implementations of predictive guidance.

IV. SUMMARY

This paper suggests that guidance law design should be considered to have two phases: determination of a policy, and synthesis of a real-time controller to implement the policy. Example techniques for both phases have been presented.

REFERENCES

1. D. M. Salmon, and P. V. Kokotovic, "Design of Linear Feedback Controllers for Nonlinear Plants", IEEE Trans. on AC, Vol. AC-14, No. 3, June 1969, pp 289-292.
2. V. Garber, "Optimum Intercept Laws for Accelerating Targets", AIAA Journal, Vol. 6, No. 11, November 1965, pp 2196-2198.
3. R. E. Dickson and V. Garber, "Optimum Rendezvous, Intercept, and Injection", AIAA Journal, Vol. 7, No. 7, July 1969, pp 1402-1403.
4. D. M. Salmon, "Predictive Guidance Policies for Intercept and Rendezvous of Non-Maneuvering Targets", Technical Note 25, Project PAU 957-006, Systems Control, Inc., Palo Alto, California, April 1971.
5. D. M. Salmon, "Strategies for Defending a Hardsite Against a Maneuvering Reentry Vehicle", Report 2, Contract DAHC60-69-C-0124, Systems Control, Inc., Palo Alto, California, April 1970 (CONFIDENTIAL).
6. R. Isaacs, Differential Games, Wiley, 1956.
7. A. E. Bryson, and Y. C. Ho, Applied Optimal Control, Blaisdell, 1969.
8. D. M. Salmon, "Policies and Controller Design for a Pursuing Vehicle", IEEE Trans. on AC, Vol. AC-14, No. 5, October 1969, pp 482-488.
9. D. M. Salmon, "Multipoint Guidance - An Efficient Implementation of Predictive Guidance", Technical Note, Project PAU 957-006, Systems Control, Inc., Palo Alto, California.

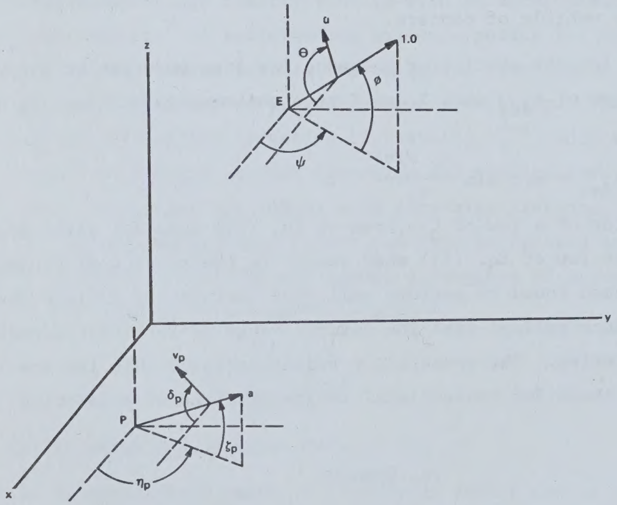


Figure 1. Variables of the Evader, E, and the Pursuer, P

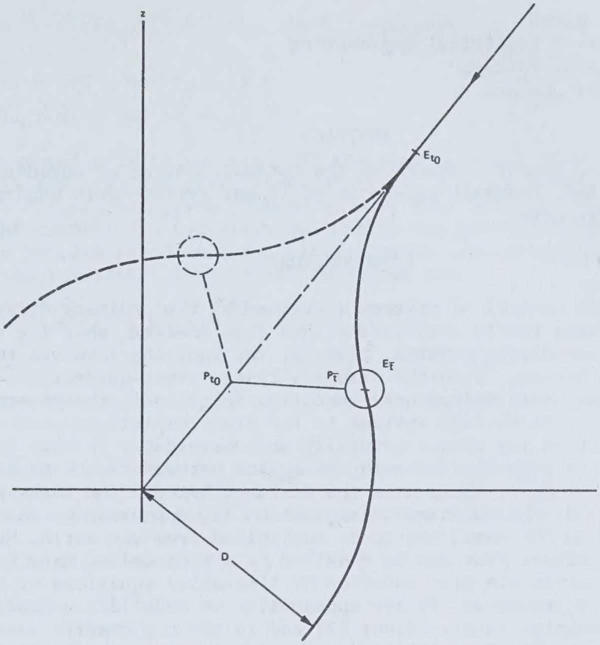

z
x
E_t0
P_t0
P_T̄
E_T̄
D

Figure 2. Optimal Trajectories for a Pursuer Able to Turn Instantaneously

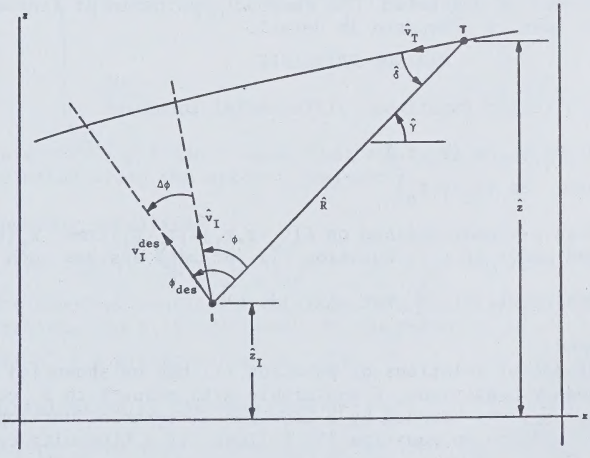

z
T
v̂_T
δ̂
γ̂
R̂
ẑ
Δφ
v̂_I
φ
v_I^des
φ_des
I
ẑ_I
x

Figure 3. Relative States of T and I

NEW TRENDS IN THE CONTROL OF HEREDITARY SYSTEMS

HEIKKI N. KOIVO
Department of Electrical Engineering
University of Toronto
Toronto 181, Canada

ABSTRACT

A survey of recent results on the optimal control of hereditary systems is presented. Feedback synthesis of linear systems with quadratic cost is considered in detail.

INTRODUCTION

The optimal control of systems described by the ordinary differential equations has been farely well established [1]. However, when the plant is modelled by a hereditary process, that is, the equation involves the past history of the process, even the simplest linear plant-quadratic cost problem was satisfactorily solved only recently. In general, the theory of optimal control for hereditary systems is far from complete.

Transportation lag occurs naturally and unavoidably in many interconnected systems. A pipe-line, serving as a link between two chemical tanks, causes a delay of $h=d/v$ where d is the distance between the tanks and v is the velocity of the substance. Information lag appears e.g. when an unmanned vehicle on the moon, has to be controlled from the earth. Human operator in the feedback link can be modelled by a time delay. Many biological growth processes are also modelled by time-delay equations or more general hereditary processes. Delays appear also in modelling a traffic network of oversaturated intersections [2] and in the air traffic control as is pointed out in [3]. More detailed reference lists appear in [4,5].

Often the lags are simply ignored in modelling. Then, if the system performance is optimized, the resulting design will only be suboptimal, because of the modelling error.

In this paper a brief survey of the results on the optimal control of hereditary systems is presented. The feedback synthesis of linear systems with quadratic cost is discussed in detail.

MAXIMUM PRINCIPLE

Consider a vector functional differential equation

$$\dot{x}(t) = F(t,x_t) \quad \text{on } [t_o,t_1] \qquad (1)$$
$$x(t) = g(t) \quad \text{on } [t_o-r,t_o]$$

where $x(t)$ is an n-vector defined on $C[t_o-r,t_1]x[t_o,t_1]$, $r>0$, $x_t(s)=x(t+s)$, $t_o-r=t+s=t$, and $g\epsilon C[t_o-r,t_o]$. Equation (1) includes systems such as

$$\dot{x}(t) = f(x(t),x(t-1), \int_{t-r}^{t} K(t,s)x(s)ds,t)$$

as special cases.

The existence of solutions of equation (1) can be shown [6] under the usual Caratheodory conditions; F measurable with respect to t, continuous with respect to x_t, and bounded by a Lebesgue integrable function. The uniqueness of the soution to equation (1) follows, if a Lipschitz condition is imposed on F.

Once the existence and uniqueness of the solution has been established the optimal control problem can be posed.

Consider the differential equation with delays

$$\dot{x}(t) = f(x(t), x(t-h_1), \ldots, x(t-h_m), u(t), u(t-k_1), \ldots, u(t-k_j), t) \quad \text{for } t > t_o \qquad (2)$$

$$x(t) = g(t) \quad \text{on } [t_o - h_m, t_o]$$

$$u(t) = v(t) \quad \text{on } [t_o - k_j, t_o)$$

where the n-vector $x(t)$ is the state of the system, the r-vector $u(t)$ is the control, $g(t)$ and $v(t)$ are given time functions, and the delays $0 < h_1 < h_2 < \ldots < h_n$, $0 < k_1 < k_2 < \ldots < k_j$ are given constants. Note that both the history of the state and the control are necessary to specify the current state.

The problem is to select a controller on time interval $[t_o, t_1]$ from a restraint set $\Omega \subset R^r$ so as to minimize the cost

$$C(u) = f^1(x(t_1), t_1) + \int_{t_o}^{t_1} f^o(s, x(s), x(s-h_1), \ldots, x(s-h_m), u(s), u(s-k_1), \ldots, u(s-k_j))\, ds \qquad (3)$$

Here the simplest point-wise form of a maximum principle is stated (no control retardation):

Theorem: Suppose that $u^*(t)$ is the optimal control and that $x^*(t)$ is the generated optimal trajectory. Then corresponding to $u^*(t)$ and $x^*(t)$, there exists a costate vector $p^*(t)$ such that the following relations hold

1) Canonical equations

$$\dot{x}^*(t) = \frac{\partial H^*}{\partial p(t)}$$

$$\dot{p}^*(t) = \begin{cases} -\left\{ \dfrac{\partial H^*}{\partial x(t)} + \displaystyle\sum_{i=1}^{m} \left[\dfrac{\partial H^*}{\partial x(s-h_i)} \right] \Bigg|_{s=t+h_i} \right\} & t_o < t < t_1 - h_m \\ -\left\{ \dfrac{\partial H^*}{\partial x(t)} + \displaystyle\sum_{i=1}^{m-m_1} \left[\dfrac{\partial H^*}{\partial x(s-h_i)} \right] \Bigg|_{s=t+h_i} \right\} & t_1 - h_{m+1-m_1} < t < t_1 - h_{m-m_1} \\ & m_1 = 1, 2, \ldots, m-1 \\ -\dfrac{\partial H^*}{\partial x(t)} & t_1 - h_1 < t < t_1 \end{cases} \qquad (4)$$

where $H = f^o + p' f$ and * means that the partial derivatives are evaluated along the optimal trajectory.

2) Boundary conditions

$$x^*(t) = g(t) \quad \text{on } [t_o - h_m, t_o]$$

The boundary conditions on p are determined as in undelayed problems and p is continuous at the points $t_1 - h_i$

$$p(t - h_i^+) = p(t - h_i^-) \quad i = i, 2, \ldots, m$$

3) Minimization of the Hamiltonian

$$H[x^*(t), x^*(t-h_1), \ldots, x^*(t-h_m), p^*(t), u^*(t)] \leq$$
$$H[x^*(t), x^*(t-h_1), \ldots, x^*(t-h_m), p^*(t), u(t)]$$

for every t, $t_o \leq t \leq t_1$, and all $u(t) \in \Omega$.

The only difference between this maximum principle and when no delays are present is in the adjoint equation. The adjoint equation turns out to be an advanced type of differential equation that has to be solved backwards.

Kharatishvili was the first one to give the pointwise maximum principle, when only the state was retarded [7]. Later he generalized the maximum principle for the problem posed above [8] by using quasi-convex families of functions. Correction to the latter proof was made in [9]. Friedman [10] considered a hereditary control problem obtaining necessary conditions. Banks [11] has obtained a maximum principle for the above problem, when the delys are time-varying and there are no control delays, in integral form under weaker conditions. Chyung and Lee use geometric arguments to prove the maximum principle for the case of linear plant (only state is retarded) and quadratic cost [12] and later generalize this to a larger class of linear functional differential equations [13]. Banks and Jacobs [14] have derived the most general maximum principle for linear functional differential equations.

Necessary and sufficient conditions for optimality have been derived in [15,16] and also in [17] for the case when the delay depends on the state and/or control. Necessary conditions for optimality were derived in [18] for neutral equations. Equation (1) is called neutral if the delayed time derivative of the state appears on the right hand side of equation (1).

LINEAR PLANT-QUADRATIC COST

The maximum principle provides the open loop control, but for design purposes a closed loop control offers many advantages. The synthesis problem can be posed as follows:

Suppose the state transition of a system is governed by

$$\dot{x}(t) = A_o x(t) + A_1 x(t-1) + B_o u(t) + B_1 u(t-a) \quad \text{for } t>t_o \tag{6}$$

$$x(t) = g(t) \quad \text{on } [t_o-1, t_o]$$

$$u(t) = v(t) \quad \text{on } [t_o-a, t_o]$$

where all the matrices are constant and of appropriate dimensions and a is a positive constant.

The cost functional to be minimized is

$$C(u) = \int_{t_o}^{T} [x'(s)Qx(s) + u'(s)Ru(s)]ds \tag{7}$$

where Q is symmetric positive semidefinite matrix, and R is a symmetric positive definite matrix.

The problem is to minimize the cost functional subject to system constraint (6) by dtermining an admissible control among the class of unconstrained controllers. Moreover, the optimal controller is to be in the feeddback form, i.e., as a function of process history x(s), $t_o \leq s \leq t$, u(s), $t_o \leq s < t$.

The problem was first considered by Krasovskii [19] when $B_1 \equiv 0$ in equation (6) and the terminal time T is infinite. He concludes from the stability results that the optimal feedback control must have the form

$$u^*(t) = - R^{-1}B_o' \left[E_o(t)x(t) + \int_{-1}^{0} E_1(t,s)x(s)ds \right] \tag{8}$$

Although the equations that $E_o(t)$ and $E_1(t,s)$ satisfy can be concluded from his paper, the presentation lacks precision and rigour. The problem has created more interest in U.S.A. at the end of 60's, when when first in [20] Krasovskii's result was given a precise proof. Then simultaneously Alekal [21] and Eller et. al. [21] utilized the Caratheodory Lemma (or equivalently the dynamic programming method) to find the optimal feedback controller when the terminal time was finite and $B_1=0$ in equation (6). It was shown by them that the coefficient matrices $E_o(t)$ and $E_1(t,s)$ satisfy the following set of equations, generalized Riccati equations, with specified boundary conditions:

$$\frac{d}{dt}E_o(t) + A_o'E_o(t) + E_o(t)A_o - E_o(t)BR^{-1}B'E_o(t) + E_1'(t,0) + E_1(t,0) + Q = 0$$

$$\frac{\partial}{\partial t}E_1(t,s) - \frac{\partial}{\partial s}E_1(t,s) + (A_o' - E_o(t)BR^{-1}B')E_1(t,s) + E_2(t,0,s) = 0$$

$$\frac{\partial}{\partial t}E_2(t,r,s) - \frac{\partial}{\partial r}E_2(t,r,s) - \frac{\partial}{\partial s}E_2(t,r,s) - E_1'(t,r)BR^{-1}B'E_1(t,s) = 0$$

$$E_o(T) = 0 \quad ; \quad E_1(T,s) = 0 \quad ; \quad E_2(T,r,s) = 0$$

$$E_1(t,-1) = E_o(t)A_1 \quad ; \quad E_2(t,-1,s) = A_1'E_1(t,s) \tag{9}$$

where $E_o(t)$ and $E_2(t,r,s)$ are positive semidefinite matrices and $E_2(t,r,s) = E_2'(t,s,r)$.

In [23] it is established that the required solution of the system of equations (9) exists and is unique. The key in finding the solution is guessing the correct form of the V-function. In this case it is

$$V(x_t,t) = x'(t)E_o(t)x(t) + x'(t)\int_{-1}^{0} E_1(t,s)x(t+s)ds$$
$$+ \int_{-1}^{0} x'(t+s)E_1'(t,s)ds\ x(t) + \int_{-1}^{0}\int_{-1}^{0} x'(t+r)E_2(t,r,s)x(t+s)drds$$

By computing

$$\frac{d}{dt}V(x_t,t) + x'(t)Qx(t) + u'(t)Ru(t)$$

and minimizing this expression with respect to u(t), the unique control given by equation (8) is obtained.

More general linear systems have been considered in the literature by adding a term $\int_{-1}^{0} A_2(t,s)x(s)ds$ to the right hand side of equation (6) $(B_1 \equiv 0)$. In [24] iteration on policy space is used under strict differentiability conditions of matrices $A_o(t)$, $A_1(t)$, $A_2(t,s)$. Transforming the same delay equation into a partial differential equation and using dynamic programming a similar result was obtained in [25]. In [26,27] time-varying delays are treated and the applicability of the Caratheodory technique is demonstrated.

It was also of interest to see if one could start from the necessary conditions of the maximum principle and derive the optimal feedback result by some similar method as is the sweep-method for the ordinary differential equations. On the surface, it appears that one cannot do this, because the state equation is delayed and the adjoint equation is advanced. However, it was shown by Manitius [28] that changing these differential equations into integral equations, it is possible to obtain feedback solution. The tool used by him was the Fredholm resolvent. The optimal feedback law obtained by him had the form

$$u^*(t) = -R^{-1}B_0'P(t,t,t) - R^{-1}B_0' \int_{t-1}^{t} P(t,s+1,t)A_1x(s)ds \qquad (11)$$

where the resolvent $P(t,s,\tau)$ satisfies the integral equation

$$P(t,s,\tau) = M(t,s) - \int_{\tau}^{T} P(t,r,\tau)BR^{-1}B'M(r,s)dr \qquad (12)$$

where

$$M(t,s) = \int_{\max[t,s]}^{T} X'(r,t)QX(r,s)dr$$

and

$$\frac{\partial}{\partial t} X(t,s) = A_0X(t,s) + A_1X(t-1,s) \qquad t_0 \leq s \leq t$$

$$X(t,t) = I \text{ (identity matrix)}$$

$$X(t,s) = 0 \text{ for } t<s$$

In [29] a similar result is established in a more abstract setting by using techniques developed by Lions for the control of partial differential equations [30].

H. Koivo [31] shows that the generalized Riccati equations and the Fredholm resolvent $P(.,.,.)$ are equivalent.

It appears that the dynamic programming fails, if $B_1 \neq 0$. This conclusion has been made in [3,25], probably, because the same V-function was used. However, if it is realized that the current state $x(t)$ depends not only on the previous history of the state $x(t)$ but also on the previous history of the control $u(t)$, then by adding the contribution due to the history of the control to the V-function, dynamic programming can be utilized to find the optimal feedback controller

$$u^*(t) = -R^{-1}B_0'[P(t,t,t)x(t) + \int_{t-1}^{t} P(t,s+1,t)A_1x(s)ds + \int_{t-a}^{t} P(t,s+a,t)B_1u(s)ds]$$
$$-R^{-1}B_1'[P(t+a,t,t) + \int_{t-1}^{t} P(t+a,s+1,t)A_1x(s)ds + \int_{t-a}^{t} P(t+a,s+a,t)B_1u(s)ds] \qquad (13)$$

where $P(.,.,.)$ satisfies a generalized matrix Riccati equation (9) with three additional shifted quadratic terms. The same result can also be derived using the Fredholm resolvent [31].

By first transforming equation (6) into a partial differential equation, discretizing this, using the standard results for discrete systems and taking the limit (formally) the same result is established in [32].

The optimal feedback law for the linear neutral equation and quadratic cost was obtained in [33]. Feedback for a class of nonlinear functional differential equations was considered in [26] and partial results were obtained.

These are the most important results for hereditary systems in the area of linear plant-quadratic cost.

CONCLUSIONS

A survey of the significant results on the control of hereditary systems was presented. The discussion concentrated on the feedback synthesis of linear hereditary systems where especially the application of the Caratheodory technique and Fredholm resolvent have resulted in elegant new results. The survey is not intended to be complete, but to indicate teh most important results in this area.

REFERENCES

1. E.B. Lee and L. Markus, Foundations of Optimal Control Theory, John Wiley and Sons, Inc., N.Y., 1967.

2. K. Burhardt,"Urban Traffic System Optimization", Ph.D. dissertation, University of Minnesota, Minneapolis, Minn., 1971.

3. H.T. Banks, M.Q. Jacobs, and M.R. Latina," The synthesis of optimal controls for linear problems with retarded controls", CDS Technical Report 70-4, Center for Dynamical Systems, Brown University,July,1970.

4. R. Bellman and J.M. Danskin," A survey of the mathematical theory of time lag, retarded control, and hereditary processes", RAND Report R-256, 1954.

5. N.H. Choksy,"Time-lag systems ... Bibliography", I.R.E. Trans. Automatic Control 5AC (1960),66-70.

6. M.N. Oǧuztöreli,"Time Lag Control Systems", Academic Press,N.Y., 1966.

7. G.L. Kharatishvili,"The maximum principle in the theory of optimal processes with delay"; 136(1961), USSR, 39-42.

8. G.L. Kharatishvili," A maximum principle in extremal problems with delays", in Matehematical Theory of Control (Eds. Balakrishnan and Neustadt), Academic Press, N.Y.,1967.

9. A. Wierzbicki,"A maximum principle for processes with nontrivial delay of the control", Automation and Remote Control,31(1970),1543-1549.

10. A. Friedman," Optimal control of hereditary processes", Arch. Rational Mech. Anal.,15(1964),396-416.

11. H.T. Banks,"Necessary conditions for control problems with variable lags", SIAM J. on Control, 6(1968),9-47

12. D.H. Chyung and E.B. Lee," Linear OPtimal control problems with delays" , SIAM J. on Control,4(1966), 548-575.

13. D.H. Chyung and E.B. Lee,"Delayed action control problems", Automatica, 6(1970),395-400.

14. H.T. Banks and M.Q. Jacobs,"The optimization of trajectories of linear functional differential equations", SIAM J. on Control, 8(1970).

15. R. Gabasov and S.V. Churakova,"Necessary conditions for optimality in systems with a delay", Automation and Remote Control, No.1 (1968), 37-54

16. R. Gabasov and S.V. Churakova,"Sufficient conditions for optimality in systems with a delay", Automation and Remote Control, No.1(1968), 193-209.

17. W.H. Ray and M.A. Soliman,"The optimal control of processes containing pure time delays- I. Necessary conditions for the optimum", Chemical Eng. Sci., 25(1970), 1911-1925.

18. G.A. Kamenskii and E.A. Khvilon,"Necessary condition for the optimal control of systems with deviating argument of neutral type",Automation and Remote Control, No. 3(1969),327-339.

19. N.N. Krasovskii,"On the analytic construction of an optimum control in a system with time lags", J. Appl. Math. Mech., 26(1962), 50-67.

20. D.W. Ross and I. Flügge-Lotz,"An optimal control problem for systems with differential-difference equation dynamics", SIAM J. on Control, 7(1969),609-623.

21. Y. Alekal,"Synthesis of feedback controllers for systems with time delay", Ph.D. dissertation, University of Minnesota,Minneapolis, Minnesota, 1969.

22. D.H. Eller, J.K. Aggarwal, and H.T. Banks,"OPtimal control of linear time-delay systems", IEEE Trans. Automatic Control AC-14(1969),678-687.

23. Y. Alekal, P. Brunovsky, and E.B. Lee,"Synthesis of feedback controllers for time delay systems", Proc. of the All Union Conf. on Minimal Surfaces and Optimal Control, Tbilissi, USSR, september 1969.

24. M.J. Kushner and D.I. Barnea,"On the control of linear functional-differential equation and quadratic cost", SIAM J. on Control, 8(1970), 257-282.

25. A.J. Koivo and H.N. Koivo,"On dynamic programming in optimizing systems with time delay", Preprints, IFAC Kyoto Symposium on Systems Engineering Approach to Computer Control, Kyoto, Japan, August,1970.

26. C.E. Mueller,"Optimal feedback control of hereditary processes", Ph.D. dissertation, University of Minnesota,Minneaopolis, Minnesota, 1971.

27. D.H. Eller and J.K. Aggarwal,"Optimization of functional differential systems", JOTA (to appear).

28. A. Manitius,"Optimum control of linear time lag processes with quadratic performance indexes", The Fourth IFAC Congress, Warsaw, Poland, June, 1969.

29. M.C. Delfour and S.K. Mitter,"Optimal control of affine hereditary differential systems with a quadratic cost", Technical Report CRM-94, Centre de Recherches Mathematiques, Universite de Montreal, April,1971.

30. J.L. Lions,"Optimal Control of Systems Governed by Partial Differential Equations",Springer Verlag, N.Y., 1971.

31. H.N. Koivo,"Fredholm resolvent in the optimization of linear systems with state and control retardations", Ph.D. dissertation, University of Minnesota, Minneapolis, Minnesota, 1971.

32. M.A. Soliman and W.H. Ray,"Optimal Feedback control for linear-quadratic systems having time delays", to be published in Int. J. Control.

33. D.H. Eller and J.K. Aggarwal,"Optimal control of neutral systems" , to be published in Int. J. Control.

RESOLUTION SPACE - A FUNCTION ANALYTIC SETTING FOR CONTROL THEORY

RICHARD SAEKS
Department of Electrical Engineering
University of Notre Dame
Notre Dame, Indiana 46556

Practitioners of the art have long recognized the great potential of function analytic and operator theoretic methods in control and system theory. Unfortunately, this potential has never been fully realized because of the lack of a satisfactory time concept on the abstract spaces in which such theories are traditionally formulated. Typically, the application of function analytic optimization procedures leads to non-causal or unstable controllers, etc.

Over the past several years an attempt has been underway to alleviate this difficulty through the medium of the (Hilbert) resolution space which adds to the usual Hilbert space an appropriate time-structure. This in turn allowing one to study time based control concepts such as causality, stability, state and controllability in a function analytic setting. In the present paper this theory is reviewed from the point of view of the control engineer.

We begin in the following section with a number of examples which indicate the potential power of the functional analytic approach in control theory but which fail to yield physically realizable systems. We then introduce the (Hilbert) resolution space as an alternative setting for such studies wherein the realizability problems can be circumvented, illustrating the concept with a number of examples arising from various system theoretic contexts. This is followed by a short review of the literature in the area indicating the state of the art and the outstanding open problems, and a bibliography, the latter including (we believe) all papers specifically concerned with resolution space as well as a number of related works which employ "resolution space like" techniques but not the formalism.

I. ASHES

Consider the control system shown in Figure 1 for which we desire to choose a controller M, given the plant T and the feedforward, N, so as to minimize the functional

$$J = ||c||^2 + ||u||^2 \tag{1}$$

where the variables lie in some Hilbert space, say $L_2[0,1]$, and the input-output relationship of the components are interpreted as operators on that space. More generally, each variable may lie in a different Hilbert space whence the components become operators mapping one Hilbert space to another. Now, if we interpret r as an initial condition (in E^n), N is the operator which maps this initial condition into the natural response (in $L_2[0,\infty]$) it engenders when applied to a given system, and T as the input-output mapping for the system (with zero initial condition); then this optimization problem is precisely the classical regulator problem. Of course, the initial condition space need not be finite dimensional as might be the case in a distributed parameters system. For instance, in a system of transmission lines the state space may be taken to be an infinite dimensional Sobelev space. Alternatively, the configuration of Figure 1 may be interpreted as a servo-mechanism, a differential game, etc. by appropriately reinterpreting the various operators and Hilbert spaces.

Consistent with the above the optimal control problem in Figure 1 is, in some sense, a universal configuration for optimal control. What is, however, more surprising is that with a relatively small amount of algebraic manipulation the problem can be "solved" explicitly[12]. In fact, the mimimizing controller is characterized by

$$M = -\frac{1}{2}(1 + T^*T)^{-1}T^*N \qquad (2)$$

While the controller of equation 2 is the unique minimizing controller it however, may not represent a causal operator even when both T and N are causal (as of course, are all real world devices). For instance, if N is the identity and T is a unit delay then M will be (minus) a unit predictor, and hence unrealizable[12].

From the above example it is clear that we have improperly formulated our optimal control problem. Rather than asking for an arbitrary controller which minimizes the functional we should be looking for the causal controller which minimizes the functional (over the class of all causal controllers). Unfortunately, our new problem cannot be satisfactorily formulated in the Hilbert space setting since the latter does not admit a well-defined time concept. Of course, most of our common examples of Hilbert Space, L_2, l_2, and l_2^+, etc. do have a natural time concept; this, however, is not included within the Hilbert space axioms and hence the theory of operators on a Hilbert space is oblivious to the time-structure when it exists. Rather, we must formulate a new space, the (Hilbert) resolution space, which has the usual properties of a Hilbert space and in addition an appropriately defined time concept, wherein the causal optimization problem is well-defined.

Although the above control problem has motivated many of the researchers in the area of resolution space we could equally well have considered a problem from stochastic filtering theory[6,8] or network synthesis[14] and reached the same conclusion - IN ORDER TO STUDY THE PROBLEMS OF SYSTEM THEORY IN A FUNCTION ANALYTIC SETTING IT IS NECESSARY TO INTRODUCE A TIME CONCEPT INTO THE SPACE STUDIED.

II. PHOENIX

Since the main difficulty with the classical Hilbert space approach to control theory is the non-realizability of the controllers derived therefrom, it is natural to begin our formulation of the resolution space approach via a study of the classical concept of causality. In the space $L_2(-\infty,\infty)$ an operator is said to be causal if whenever one is given two functions, f and g such that

$$f(t) = g(t);\ t \leq r \qquad (3)$$

for some r[Δ] then

$$[Tf](t) = [Tg](t);\ t \leq r \qquad (4)$$

Of course, if one assumes linearity and/or time-invariance this definition may be greatly simplified, whereas if t and r are required to be integers (positive integers) the same definition holds for the sequence space l_2 (l_2^+).

Since we desire, ultimately to extend the definition of causality to an arbitrary Hilbert space we desire to reformulate it in such a manner that the function space characteristics of the space are not required, at least explicitly. For this purpose we define the family of truncation operators E^r via

$$[E^r f](t) = \begin{cases} f(t); & t \leq r \\ 0 & ; t < r \end{cases}$$

[Δ] Equalities in L_2 are of course only required to hold almost everywhere. For brevity, however, we drop the explicit statement of this fact.

in which case the definition of causality for an operator, T, on L_2 may be reformulated by requiring that $E^rTf = E^rTg$ whenever $E^rf = E^rg$. Now a little algebra will reveal that this definition of causality is identical to the classical one, and also like the classical definition it remains valid in essentially any function or sequence space once the obvious modifications are made. Unlike the classical definition, however, the fact that L_2 is a function space only appears in the definition of E^r, not in the definition of causality. As such if one could delineate the properties of the family of operators, E^r, and define such axiomatically one could formulate a well-defined causality concept in an abstract space. This is, in fact, possible with the E^r taking the form of a resolution of the identity in an arbitrary Hilbert space H. That is:

i) E^r is a projection ($E^r = E^{r*} = E^rE^r$) for each r.

ii) E^r is increasing (Range $\{E^r\} \sqsubset$ Range $\{E^s\}$ if $r \leq s$).

iii) $\underset{r \to \infty}{\text{s-limit}}\, E^r = I$ and $\underset{r \to \infty}{\text{s-limit}}\, E^r = 0$.

iv) E^r is a (strongly) continuous function.

where the time parameters r lie in an ordered (LCA) group, G; for instance, the reals or the integers. We thus define a (HILBERT) RESOLUTION SPACE AS A PAIR (H,E^r) WHERE H IS A HILBERT SPACE AND E^r IS A RESOLUTION OF THE IDENTITY IN H.

Of course, our classical fucntion space concepts are included within the above formalism simply by taking E^r to be the family of truncation operators on L_2, l_2, l_2^+, etc. More generally, however, one can define a number of alternative non-classical causality concepts via the resolution space approach some of which do, indeed, prove to be physically relevant. For this purpose we interpret E^rx as that part of x before time r and $(I-E^r)x \triangleq E_rx$ as that part of x after r. Of course, this is just the natural interpretation when x is a function and E^r the family of truncation operators. It, however, also yields viable concepts of "past" and "future" when the E^r are not truncation operators on a function space. For instance, in a problem where the signals obey some specified statistical criterion one might take as the past the measured signal before time r and our best estimate of the signal after r based on the measured signal before r. Of course, in this case the future may be interpreted as the error between the best estimate of the signal after r and the actual signal after r. This formulation of the concepts of past and future arises naturally in certain filtering problems[8] falls readily into our abstract setting. Here we define our resolution of the identity by

$$[E^rf](t) = \begin{cases} f(t); & t \leq r \\ \hat{f}(t); & t \overline{<} r \end{cases} \tag{6}$$

where $\hat{f}$ is the best (linear) estimate of f given f(t) $t \leq r$ based on the specified statistics. Such problems are usually studied in a reproducing kernel Hilbert space wherein, depending on the statistics, we may have $\hat{f}(t) = f(r)$ $t > r$, $\hat{f}(t) = f(r)e^{-\sigma(t-r)}$ $t > r$, etc. With this choice of E^r we obtain a perfectly viable resolution space for which time based concepts such as causality are well defined and possess the same kind of properties as the classical causality concept, though predicated on a different definition of past and future.

In essence the (Hilbert) resolution space serves as a single abstract setting in which the classical time concepts of mathematical system theory can be unified and extended. Unlike an arbitrary Hilbert space, however, the difficulties indicated in the previous section in applying function analytic methods to control problems do not arise when a resolution space setting is employed. Not only it is possible to handle causality questions in such a setting but also problems associated with the concepts of state, stability, passivity, time-invariance, Fourier representation, etc.

Of the various types of operator which are well-defined on a resolution space wherein time is a meaningful concept but not in an arbitrary Hilbert space the most important are the

Causal Operators. For this class of operators which is of fundamental importance in the optimal control problems considered in the introduction as well as most problems arising in mathematical system theory wherein a realizability constraint must be invoked the basic properties have been formulated in both the linear[10,13,14] and nonlinear[4,11,13] cases. Various alternative modes of causality such as strict causality, anti-causality, etc. have been defined[4,14] and their representation as triangular operators[4,14] relative to an appropriate integral representation[14] obtained. Of particular interest to control theory is the decomposition of a given operator into the sum or product of a causal operator and an anti-causal operator. In the case of the multiplicative decomposition a rather complete theory has been developed for the general case[14] and various special cases[12,13,16]. On the other hand, a number of rather deep problems concerning the additive decomposition remain open[4]. Precise condition for the existence of an additive decomposition remains unknown though a characterization of the decomposition and a proof of uniqueness for the case when the decomposition exists has been given[14] along with the formulation of a number of special cases wherein existence of the decomposition is assured[4,17]. Another open problem concerning the causal operators is the delineation of conditions under which the inverse of a causal operator will be causal. Here only sufficient conditions are known[4,5,13,14].

Closely related to the causal operators are a number of classes of

Operators Which Satisfy Energy Constraints. These operators typically are required to satisfy some boundedness or positivity constraint formulated in terms of the truncated inner product $^{t}<x,y> = <E^{t}x,E^{t}y>$, the particular definition depending on the physical interpretation of the variables. Various classes of operators which have been studied[10,11,13,14,20] include passive, lossless, disipative, non-energetic and energy limited operators, the relationship between these concepts and causality having been formulated both in the linear[10,14] and nonlinear[11,13] cases. Typically, an operator is shown to satisfy an energy constraint if and only it if is causal and satisfies some similar energy constraint formulated in terms of the usual inner product. Since the energy constraints are formulated in terms of the truncated inner product the resultant classes of operator are, like the causal operators, well-defined in a resolution space but not in the usual Hilbert space context. A number similar operator classes of this type which have yet to be formalized but may prove to be of considerable importance are entropy preserving or decreasing operators and operators with special types of energy storage capability (for instnace, electric or magnetic energy).

In essence resolution space differs from the classical spaces usually associated with operator theory in that it has a well-defined time structure thereby allowing concepts which are time dependent to be formulated. One such concept which is of considerable importance in control theory is the State of an operator. Conceptually, the state of an operator at time t is a parameter which characterizes how the input before t effects the output after t. As such, given an operator T, on a resolution space one is interested in the "reduced" operator $E_{t}TE^{t}$ where $E_{t} = I - E^{t}$ is the operator which picks out that part of the response after t and E^{t} is the usual "truncation" which picks out that part of the input before t. An appropriate factorization of this family of operator has been shown to yield a viable state-space concept which exhibits most of the fundamental properties associated with the classical state concept for differential and difference equations yet is applicable to essentially any linear[16] or nonlinear[4] operator. Given this formulation of the state concept in resolution space various related topics such as controllability and observability[16,19,21], realization, minimality[16] and stability[16,21] may be studied. Remaining open problems include the extension of the aforementioned studies to the nonlinear case via the recently formulated nonlinear decomposition theory[4] and the development of an appropriate "differential" representation theory for operators with a specified state decomposition.

Consistent with the classical approaches to control theory one would like to obtain a frequency domain theory along with the time domain theory for operators and systems defined on a resolution space; hence a major area of study is concerned with

Fourier Analysis. Unlike the time domain studies thus far described if one is to formulate a viable frequency domain concept the given resolution space must admit an appropriate "dual," the time parameter in the original space being replaced by a frequency parameter in the "dual." Although the appropriate class of resolution spaces, termed uniform[16,17] and semiuniform[18] resolution spaces is somewhat restricted it does include most of the spaces classically encountered in control theory along with many interesting generalizations thereof. Unlike the classical frequency domain formulation wherein one takes the Fourier transform of the input and output of a system and then formulates the "frequency response" of the system so as to relate the corresponding Fourier transforms, in the resolution space context the inputs and outputs may not explicitly be identified as function; hence we construct the frequency response directly from the operator representing the system without first transforming the inputs and outputs [17]. The resultant frequency response can be defined via an appropriately generalized spectral decomposition of the given operator[17] or via a Fourier transform[6] or Fourier-Stieltjes[17] transform of an appropriate time-domain representation of the given system. The frequency domain formulation via resolution space is quite new and has many open problems. In particular, the Fourier and Fourier-Stieltjes transform formulation remains less general than the spectral theoretic formulation, operator valued distributional techniques being required to match the two theories. The Fourier representation of time-variable systems, a topic which has only been partially developed even in the classical function spaces, remains untouched in the resolution space formulation though the latter has the potential of yielding results which are not readily apparent via the classical formulation. Finally, system function, or Laplace representation, techniques have only been touched upon thus far[17], the concept having been defined only for a restricted class of systems which also admit a well-behaved frequency response though one would expect that the Laplace formulation should be more - not less - general than the Fourier formulation.

Of the various applied topics in control theory which have been studied in a resolution space setting the one which has received, by far, the greatest attention and is thus the most fully developed is

Stability. These studies fall essentially into two categories - input-output stability of feedback system[1,2,3,5,20] and stability of a state-space decomposition[15,19,21]. In the former case the "classical" extension space approach has been shown to be equivalent to a requirement that return difference for the system satisfy a number of axioms formulated in a resolution space (invertibility, causality, etc.)[20] these in turn having been translated into sufficient conditions for stability in both the linear[5] and nonlinear[1,2,3,20] cases. In particular, various strict causality constraints on the plant lead to stable feedback systems[5]. The theory for the stability of a state space representation is somewhat less fully developed most work having been directed at exhibiting the relationship between internal and external stability given appropriate controllability and observability assumptions. Both the linear[15] and nonlinear[21] cases have been studied though considerable work remains to be done. Open problems include further development of the state space stability conditions, continued exploitation of the strict causality concept and the development of frequency domain stability conditions.

Since one of the main motivations for the development of resolution space was the solution of the realizability problems associated with the various function analytic approaches to the problems of

Optimal Control, this area has received considerable attention. The problem of main interest has been the control problem described earlier for which optimal causal solutions have been constructed in various special cases[12,13] the sensitivity and stability of such having been studied[13]. The problem of optimal control via state feedback, on the other hand, has receceved considerably less attention. A theorem verifying the sufficiency of memoryless feedback for this case has been formulated[15] but a formal theory for obtaining the feedback operator has not yet been developed. Open problems in the area thus remain numerous - most of the state space optimal control problems, further development of input-output optimal control and the development of frequency domain methods for optimal control.

A number of other areas of mathematical system theory which have received attention in the resolution space literature but for which the surface has barely been scratched include

Filtering, Network Synthesis, Sensitivity and Stochastic Control. In the case of filtering theory, primary concern has been centered around the problem of whitening the signal statistics by appropriate choice of the resolution space[6,9]. That is, if one chooses the appropriate resolution space it is possible to force a given set of statistics to have the characteristics of white noise (from the point of view of that space) even though they may not be white in the function space where they would normally be defined. As such, it is possible to employ techniques traditionally associated with white noise to filtering problems characterised by essentially arbitrary statistics though in a "non-standard" resolution space. Of course, such techniques are also applicable to stochastic control though they have yet to be exploited. In network theory the concepts of passivity and losslessness have been formalized[10,11,14] and the problems of active and passive network synthesis studied with existence theorems being formulated. The lossless synthesis remains open as does the majority of the theory. Finally, the sensitivity characteristics of optimal controllers have been formulated [12,13] from a sensitivity matrix point of view, the properties thereof being related to the various energy constraints[13]. At this time no general sensitivity measure has been developed nor has the problem of synthesizing minimal sensitivity control systems been studied.

Of a more abstract nature than the aforementioned studies is the development of a

Representation Theory for the resolution spaces. Here a complete set of invariants (for the category of resolution spaces together with causal maps) has been developed by appropriately adopting the classical theory of spectral multiplicity and the class of uniform resolution spaces to which the Fourier Analysis techniques are applicable has been characterized[16]. Remaining open problems include the characterization of those similarity transformations on a resolution space which preserve causality and/or time-invariance and in particular the application of such transformations to the problems of digital simulation and digital filtering. In this respect it has been conjectured that there exists a transformation mapping l_2 to L_2 which takes causal operators into causal operators if and only if the operators are piecewise constant, though a proof has only been given in the memoryless special case[9].

Finally, in the light of further generalizing the techniques of control theory and relating it to other areas of endeavor one is interested in formulating and studying various

Non-Hilbert Resolution Spaces. Of course, the most obvious generalization of the (Hilbert) resolution space is the (Banach) resolution space for which the basic properties have been formulated[13], though the status of the theory is well behind that for the Hilbert space case. Also of considerable interest is the group resolution space[4] though the work in the area has been concentrated almost exclusively on the additive decomposition problem for causal operators. Clearly, one can formulate a resolution space concept in any topological vector space or in fact even in a manifold setting though at this writing these areas remain virgin, as is the formulation of an abstract resolution concept in a category theoretic setting.

IV. FLAME

In the preceeding we have attempted to illustrate the difficulties in applying Hilbert space methods to the problems of control and to indicate one means of circumventing these difficulties - the (Hilbert) resolution space. Although far from complete at this time we believe that the results thus far obtained via this approach indicate the potentiality of obtaining a completely unified approach to control and system theory through the medium of the resolution space which would simultaneously include discrete, continuous, distributed and time-variable systems within the same theory.

V. LACTANIUS

1. Damborg, M.J., "Stability of the Basic Nonlinear Operator Feedback System," Tech. Rpt. 37, Systems Engrg. Lab., Univ. of Mich., Ann Arbor, Mich., April 1969.

2. Damborg, M.J., "The Use of Normed Linear Spaces for Feedback Systems Stability," Proc. of the 14th Midwest Symp. on Circuit Theory, Univ. of Denver, 1971, paper 7-2.

3. Damborg, M.J. and A.W. Naylor, "Fundamental Structure of Input-Output Stability for Feedback Systems," IEEE Trans. on Sys. Sci. and Cyb., Vol. SSC-6, pp. 92-96, (1970).

4. DeSantis, R., Ph.D. Thesis, Univ. of Michigan, Ann Arbor, (in preparation).

5. DeSantis, R., "Causality and Stability in Resolution Space," Proc. of the 14th Midwest Symp. on Circuit Theory, Univ. of Denver, 1971, Paper 4-3.

6. Duttweiler, D.L., Ph.D. Thesis, Stanford Univ., 1970.

7. Falb, P.L. and M.I. Freedman, "A Generalized Transform Theory for Causal Operators," SIAM J. on Cont., Vol. 7, pp. 452-471, (1969).

8. Kailath, T. and D.L. Duttweiler, "Generalized Innovation Processes and Some Applications," Proc. of the 14th Midwest Symp. on Circuit Theory,

Univ. of Denver, 1971, Paper 4-2.

9. Kaplan, W., Ph.D. Thesis, Univ. of Maryland, College Park., Md., 1970.

10. Porter, W.A., "Some Circuit Theory Concepts Revisited," Int. J. on Cont., Vol. 12, pp. 433-448, (1970).

11. Porter, W.A., "Non-Linear Systems in Hilbert Space," Int. J. on Cont., Vol. 13, pp. 593-602, (1971).

12. Porter, W.A., "A Basic Optimization Problem in Linear Systems," Math. Systems Theory, Vol. 5, pp. 20-44, (1971).

13. Porter, W.A. and C.L. Zahm, "Basic Concepts in System Theory," Tech. Rpt. 33, Systems Engrg. Lab., Univ. of Mich., Ann Arbor, 1969.

14. Saeks, R., "Causality in Hilbert Space," SIAM Review, Vol. 12, pp. 357-383, (1970).

15. Saeks, R., "State in Hilbert Space," SIAM Rev. (to Appear), Also, Tech. Memo. EE 6912-a, Univ. of Notre Dame, Notre Dame, Ind., 1969.

16. Saeks, R., "Causal Factorization, Shift Operators and the Spectral Multiplicity Function," Tech. Memo. EE-715, Univ. of Notre Dame, Notre Dame, Indiana, 1970.

17. Saeks, R., "Fourier Analysis in Hilbert Space," Tech. Memo. EE-7105, Univ. of Notre Dame, Notre Dame, Ind., 1971.

18. Saeks, R., and R.J. Leake, "On Semi-Uniform Resolution Space," Proc. of the 14th Midwest Symp. on Circuit Theory, Univ. of Denver, 1971, Paper 4-5.

19. Schnure, W.K., "Controllability and Observability Criteria for Operators on a Hilbert Resolution Space," Proc. of the 14th Midwest Symp. on Circuit Theory, Univ. of Denver, 1971, Paper 4-4.

20. Willems, J.C., "Stability, Instability, Invertibility and Causality," SIAM J. on Contr., Vol. 7, pp. 645-671, (1969).

21. Willems, J.C., "The Generation of Lyapunov Functions for Input-Output Stable Systems," SIAM J. on Cont., Vol. 9, pp. 105-134, (1971).

VI. ACKNOWLEDGMENT

The author wishes to express his thanks to R. DeSantis and W. Schnure for helping to compile the list of open problems used in preparing this manuscript.

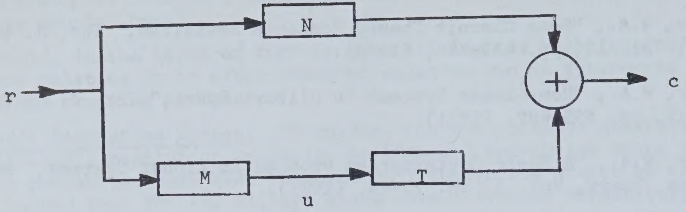


Figure 1. Open Loop Control System

GENERATION OF REGULAR EXPRESSIONS FOR AUTOMATA BY THE INTEGRAL OF REGULAR EXPRESSIONS

L. W. SMITH and S. S. YAU
Departments of Computer Sciences and Electrical Engineering
Northwestern University
Evanston, Illinois 60201

ABSTRACT

In this paper, the integral of regular expressions is defined and its properties are presented. The concept of the integral is then applied to establishing an algorithm for generating the regular expression of an arbitrary finite automaton specified by its state diagram or flow table. The regular expression for a given automaton obtained by this algorithm is shown to be unique up to the commutative property of the sum. This algorithm is effective and suitable for machine implementation. Programmed results of this algorithm are also given and interpreted.

INTRODUCTION

The relationship between regular expressions and finite automata is well known [1,2] and a number of authors [3-6] have studied the conversion between regular expressions and finite automata. McNaughton and Yamada [3] presented a pair of algorithms which systematically converted a regular expression to a finite automaton and vice versa. Although these algorithms are systematic, they have been shown to be often lengthy and too tedious. Using the techniques suggested by Arden [4], Brzozowski and McCluskey [5] gave a signal flow graph method for finding regular expressions from a finite automaton. Even though quite effective, the technique largely depends on the intuition and insight of the user. Later, Brzozowski [6] defined the derivative of a regular expression, and showed this to be the natural and most systematic way of going from a regular expression to the state diagram of an automaton.

In this paper, we will describe the notion of the integral of a regular expression analogous to that of a derivative. We will show how from a given deterministic state diagram we can use the notion of an integral of a regular expression to determine its corresponding regular expression. The developed algorithm will be shown to be systematic and effective. Finally, we will present and interpret some of the results which were obtained when the algorithm was programmed on the IBM 360/67 using SNOBOL4.

We will concern ourselves with the usual model of a finite deterministic automaton

$$M = \{A, Z, Q, f, g\} ,$$

where

$A = \{a_0, a_1, \ldots, a_{v-1}\}$ is a finite set of inputs to M.
$Z = \{z_0, z_1, \ldots, z_{w-1}\}$ is a finite set of outputs from M.
$Q = \{q_0, q_1, \ldots, q_{u-1}\}$ is a finite set of states of M.
$f\colon\ Q \times A \to Q$ is the next state function of M.
$g\colon\ Q \to Z$ is the output function of M.

The results will be presented in terms of the Moore model. The output of the automaton is the output of the final state. In the case where the output alphabet is $\{0,1\}$, the output of "1" will be denoted by a broken circle around the output state. The results also apply equally well to the Mealy model with a slight modification of the algorithm.

REGULAR EXPRESSIONS

Definition 1: A sequence over the finite alphabet A is defined to be a series of concatenated symbols from the alphabet A, λ, and $\emptyset$.

The regular operations sum, product, and star on sets of sequences P, P_1, P_2, ... over the alphabet A are defined as follows:

Definition 2.1: The regular operation sum on the sets of sequences P_1 and P_2, denoted by $(P_1 + P_2)$, is defined to be the set of sequences

$$P_1 + P_2 = \{p \mid p \in P_1 \text{ or } P_2\}. \tag{1}$$

Definition 2.2: The regular operation product on the sets of sequences P_1 and P_2, denoted by (P_1P_2), is defined to be the set of sequences

$$P_1P_2 = \{p \mid p = p_1p_2;\ p_1 \in P_1,\ p_2 \in P_2\} \tag{2}$$

Definition 2.3: The regular operation star on the set of sequences P, denoted by P*, is defined to be the set of sequences

$$P^* = \sum_{n=0}^{\infty} P^n \tag{3}$$

where $P^0 = \lambda$, $P^1 = P$, $P^2 = PP$, ...

These regular operations have the following properties:

Property 1: Sum is associative and communicative

Property 2: Product is associative but not communicative.

Property 3: Product is distributive over sum, but sum is not distributive over product.

Property 4: λ serves as the multiplicative unity.

Property 5: $\emptyset$ serves as the additive and multiplicative zero.

Now a regular expression can be defined as follows:

Definition 3: A regular expression over an alphabet $A = \{a_0, a_1, \ldots, a_{v-1}\}$, can be defined as follows:

(1) The letters a_0, a_1, ..., a_{v-1}, λ and $\emptyset$ are regular expressions.

(2) If P_1 and P_2 are regular expressions, then so are $P_1 + P_2$, P_1P_2 and P_1^*.

(3) Nothing else is a regular expression unless it follows from repeated applications of (1) and (2).

THE INTEGRAL OF A REGULAR EXPRESSION

Before we define the integral of a regular expression, we must define another operation on the set of sequences R called the derivative of R [6].

Definition 4: Given a set of sequences R and a sequence s, the derivative of R with respect to s, denoted by D_sR, is defined to be

$$D_sR = \{p \mid sp \in R\} . \tag{4}$$

We now define the operation on the set of sequences I, called the integral of I, in terms of the derivative of R.

Definition 5: Given a set of sequences I which is a derivative of the regular set of sequences R and a letter a_i, the integral of I with respect to a_i, denoted by $\int I \, da_i$, is defined to be

$$\int I \, da_i = \{I' \mid I' = a_iI,\ I' \subset R\} . \tag{5}$$

The integral of I with respect to a finite sequence $s_n = a_{i_1} a_{i_2} \ldots a_{i_n}$ is defined recursively to be

$$\int \cdots \left[\int \left(\int I \, da_{i_n} \right) da_{i_{n-1}} \right] \cdots \, da_{i_1}$$

$$= a_{i_1} \cdots a_{i_{n-1}} a_{i_n} I \subset R. \tag{6}$$

The integration of and the integration by the special letters λ and $\emptyset$ are defined as follows:

Definition 6: Given a set of sequences I, the integral of I with respect to λ is defined to be

$$\int I \, d\lambda = I \tag{7}$$

and the integral of I with respect to $\emptyset$ is defined to be

$$\int I \, d\emptyset = \emptyset \, . \tag{8}$$

Definition 7: Given the special letters λ and $\emptyset$, the integral of λ with respect to the letter a_i is defined to be

$$\int \lambda \, da_i = a_i \tag{9}$$

and the integral of $\emptyset$ with respect to a letter a_i is defined to be

$$\int \emptyset \, da_i = \emptyset \, . \tag{10}$$

Since the integrations of I with respect to λ and $\emptyset$ produces I and $\emptyset$ respectively, these integrations will be considered as trivial. We will, hence, consider only the integration of a regular expression I with respect to a series of single letters to form another integral I' which is a subset of the original regular expression R. Therefore, we will next define what we mean by the complete integral of a regular expression with respect to the alphabet A. Before we do this, we must define another integral called the λ-integral.

Definition 8: The λ-integral, denoted by Z, for the complete integral $D_s R$ is defined to be

$$Z = \begin{cases} \lambda \text{ if the sequence } s \in R \\ \emptyset \text{ if the sequence } s \notin R \end{cases} \tag{11}$$

Definition 9: Given the regular expression $D_{a_i} R$, one for every $a_i \in A$, the complete integral over the alphabet A is defined to be the regular expression

$$R = D_\lambda R = \sum_{i=0}^{v-1} \int D_{a_i} R \, da_i + Z. \tag{12}$$

If we substitute any finite sequence s of letters of the alphabet A for λ in the above equation, the complete integral $D_s R$ over the alphabet A becomes

$$D_s R = \sum_{i=0}^{v-1} \int D_{sa_i} R \, da_i + Z. \tag{13}$$

Next, consider the relationship between the set of input sequences and the states of the automaton M. Define each state q_i in terms of the set of sequences which when started in this state will carry M into an output state. We say that each regular expression defines its state q_i in terms of the output states. From this set of input sequences, we form a regular expression and obtain the following properties.

Property 6: Given an automaton M, each state q_i can be characterized by a regular expression in which each input sequence takes M from q_i into some output state.

Property 7: If two regular expressions are equivalent, then their corresponding states of the automaton are equivalent.

Let s_i and s_j be the input sequences taking the automaton M from q_0 to q_i and q_j respectively, and let R be the regular expression defining M. Then, we have the following theorem.

Theorem 1: The two states q_i and q_j of an automaton M characterized by the complete integrals $D_{s_i} R$ and $D_{s_j} R$ are equivalent if and only if the complete integrals are equivalent.

Proof: If q_i and q_j are equivalent, then every state entered from q_i via a_i must be entered from q_j via a_i and vice versa. That is $D_{s_i a_i} R$ is equal to $D_{s_j a_i} R$ for all $a_i \in A$. Also, if q_i is equivalent to q_j, both q_i and q_j have or do not have an output. Therefore,

$$\sum_{i=0}^{v-1} \int D_{s_i a_i} R \, da_i + Z = \sum_{i=0}^{v-1} \int D_{s_j a_i} R \, da_i + Z \tag{14}$$

$$D_{s_i} R = D_{s_j} R.$$

Likewise, if $D_{s_i} R = D_{s_j} R$, then $\int D_{s_i a_i} R \, da_i$ is equal to $\int D_{s_j a_i} R \, da_i$ and the state associated with $D_{s_j} R$, and $\int D_{s_j a_i} R \, da_i$ is equal to $\int D_{s_i a_i} R \, da_i$ and the state associated with $D_{s_i} R$. That is, the state entered from q_i by input letter a_i is equivalent to the state entered by the same input letter from q_j. Also Z in $D_{s_i} R$ and $D_{s_j} R$ are equivalent. Therefore, q_i is equivalent to q_j. Q.E.D.

As a consequence of this theorem, if we are given a state q_1 characterized by the regular expression R_i which is unknown and some regular expression $s_1 R_i + s_2$, derived through integration, we have

$$R_i = s_1 R_i + s_2 = s_1{}^* s_2 \, . \tag{15}$$

In order to reduce the amount of unnecessary confusion, the state q_i and the characterizing regular expression R_i will be represented by just q_i from here on, i.e. (15) can be written as

$$q_i = s_1 q_i + s_2 = s_1{}^* s_2 \, . \tag{16}$$

ALGORITHM FOR CONSTRUCTING A REGULAR EXPRESSION FROM A STATE DIAGRAM

Definition 10. Two input sequences over the alphabet A are similar with respect to an automaton M if and only if starting in the initial state of M they terminate in the same state q_i of M.

From this definition we formulate the following theorem.

Theorem 2: Given the set T of all possible finite input sequences over the input alphabet $A \cup \{\lambda\}$ and the state flow table, T can be partitioned into n similar classes, where n is the number of states in the flow table.

Proof: We know that under a deterministic system the input sequence λ carries any state back into itself. Also, we know that every finite input sequence from the initial state must end with one of the states of M.

Q.E.D.

Following directly from Theorem 2, we have the following theorem.

Theorem 3: Associated with each similar class formed from the set T of all possible input sequences over the input alphabet $A \cup \{\lambda\}$ is a state q_i.

From Brzozowski's work [4], we know that each state of an automaton M can be represented by a regular expression $D_s R$ which is a derivative of the initial regular expression R with respect to some input sequence s. Using this idea we choose the minimal input sequence (i.e. input sequence of least lexographical ordering) from each of the similar classes and form a minimal derivative $D_{s_j} R = q_j$ which is equivalent to the regular expression for the state q_j.

Definition 11: A derivative D_{s_j} for state q_j is minimal if and only if there exists no other derivative D_{s_k} for state q_j such that s_k is less lexographically than s_j.

For simplicity we will write $D_{s_j} R = q_j$, where q_j is understood to represent both the state and the regular expression representing the state. From Definition 9, we recall that each state can be represented by the complete integral

$$D_{s_j} R = \sum_{i=0}^{v-1} \int D_{s_j a_i} R \, da_i + Z, \tag{17}$$

where D_{s_j} represents some state q_j in the automaton M, and $D_{s_j a_i} R$ the state entered from q_j using the input letter a_i. Thus, we have the relation between the derivative of the regular expression used to determine the states of M and the integral of the states of M used to form the regular expression. Taking the set of minimal derivatives, which were discussed in the preceding paragraph, we place them in a lexographical list according to their input sequences. The minimal derivative associated with the null state is skipped since both its integration and differentiation will always equal $\emptyset$. Then, a tree can be constructed as follows: Starting from the first minimal derivative $D_\lambda R$ on the list, draw a branch to each $D_{a_i} R$, where $a_i \in A$. If $D_{a_i} R$ is not minimal, then no branches will be drawn from $D_{a_i} R$. If $D_{a_i} R$ is minimal, draw a branch from $D_{a_i} R$ to each $D_{a_i a_j} R$, where $a_j \in A$. Apply this process repeatedly until the list of minimal derivatives is exhausted. The corresponding state is next associated with each of these derivatives. If the derivative is equal to the null state, the corresponding state is replaced by $\emptyset$.

At this point, we have taken each state q_i of the state diagram and associated it with a derivative $D_{s_j} R$. We have shown that this derivative is also the complete integral (17). From this information we have built a tree in which each node corresponds to a complete integral, $D_{s_j} R$, and each branch corresponds to one of the integrands $D_{s_j a_i} R$ used to form it. We will now investigate four rules for manipulating these integrals based on the properties of a regular expression under integration.

Rule 1: Combinatorial Rule. If we have the following integrals over the input alphabet $\{a_i, a_j\}$

$$\int D_{sa_i} R \, da_i = a_i P_1 \quad \text{and} \quad \int D_{sa_j} R \, da_j = a_j P_2 \; ; \tag{18}$$

then we have the integral

$$D_s R = a_i P_1 + a_j P_2 \tag{19}$$

Rule 2: Output Rule. If the state represented by this integral has an output, add λ to the integral.

Rule 3: Substitution Rule. If state q_k is associated with the regular expression P_1 and the state q_j with the regular expression $P_2 q_k$, then q_j corresponds to $P_2 P_1$.

Rule 4: Star Rule. If a state q_k is associated with $P_1 q_k + P_2 + P_3$, then the integral of q_k corresponds to $P_1^*(P_2 + P_3)$.

Rules 1 and 2 are based on Definition 2.1 and Definition 8 respectively. These two rules give us the general form for the complete integral. Rule 3 is the general substitution rule and needs no further explanation. Rule 4 is a restatement of Arden's theorem [4].

Using these concepts we state the following algorithm for determining a regular expression for a given state diagram.

1. Form the flow table and lexographical list of minimal derivatives beginning with $D_\lambda R$. Go to Step 2.
2. Construct a tree from the minimal dervatives. Go to Step 3.
3. Take the branches from the node corresponding to the last minimal derivative in the list of Step 1 and integrate the derivatives of the terminal nodes of the branches. Go to Step 4.
4. Combine the separate integrals in Step 3 according to the order of the integral manipulation Rules 1 through 4 to form the complete integral of the last minimal derivative. Then, delete this last minimal derivative. Go to Step 5.
5. Repeat Steps 3 and 4 for each minimal dervative in the reverse order of the list in Step 1 until all the integration is complete.

The procedure will now be illustrated by the following example.

Step 1. Given the state diagram in Fig. 1 with q_0 and q_5 being the starting state and the null state respectively, we generate the flow table shown in Fig. 2 and the lexographical list of minimal derivatives.

$D_\lambda R$
$D_1 R$
$D_{10} R$
$D_{11} R$
$D_{110} R$

Step 2. Using the list of minimal derivatives from Step 1, the tree shown in Fig. 3 is constructed and the proper state or ∅ is associated with each entry.

Step 3. Integrating $D_{1100}{}^{R}$ and $D_{1101}{}^{R}$, we obtain

$$\int D_{1100}{}^{R}\, d0 = \int \emptyset\, d0 = 0\emptyset = \emptyset$$

$$\int D_{1101}{}^{R}\, d1 = \int q_3\, d1 = 1q_3$$

Step 4. Combining these integrals according to Rule 1, we have

$$q_4 = 1q_3$$

Step 5. Repeating Stepts 3 and 4 for other minimal derivative in the reverse order of the list of minimal derivatives in Step 1, we have

$$\int D_{110}{}^{R}\, d0 = \int 1q_3\, d0 = 01q_3$$

$$\int D_{111}{}^{R}\, d1 = \int q_3\, d1 = 1q_3$$

$$q_3 = 01q_3 + 1q_3 + \lambda \qquad \text{By Rule 1,2}$$

$$= (01 + 1)^* \qquad \text{By Rule 4}$$

$$\int D_{100}{}^{R}\, d0 = \int q_1 d0 = 0q_1$$

$$\int D_{101}{}^{R}\, d1 = \int \emptyset\, d1 = 1\emptyset = \emptyset$$

$$q_2 = 0q_1 \qquad \text{By Rule 1}$$

$$\int D_{10}{}^{R}\, d0 = \int 0q_1\, d0 = 00q_1$$

$$\int D_{11}{}^{R}\, d1 = \int (01 + 1)^*\, d1 = 1(01 + 1)^*$$

$$q_1 = 00q_1 + 1(01 + 1)^* \qquad \text{By Rule 1}$$

$$= (00)^*1(01 + 1)^* \qquad \text{By Rule 4}$$

$$\int D_0{}^{R}\, d0 = \int \emptyset\, d0 = 0\emptyset = \emptyset$$

$$\int D_1{}^{R}\, d1 = \int (00)^*1\ (01 + 1)^*\, d1 = 1(00)^*1(01 + 1)^*$$

$$q_0 = 1(00)^*1(01 + 1)^* \qquad \text{By Rule 1}$$

From the discussion of the algorithm and the illustrative example, we obtain the following basic properties for the algorithm.

Property 8: The algorithm is independent of the numbering of the states.

Property 9: The resultant regular expression for the state diagram is independent of the order in which the integration is performed.

Property 10: The resultant regular expression is dependent only on the tree formed by the minimal derivatives This tree is explicitly defined by the algorithm. These properties yield the following theorem.

Theorem 4: The regular expression for a given state diagram obtained by the above algorithm is unique up to the commutative property of the sum.

MACHINE IMPLEMENTATION

In this paper, we have considered the notion of the integral of a regular expression and have shown how it can be applied to a state diagram to generate its corresponding regular expression. The algorithm has been implemented on IBM 360/67 computer. The results of this implementation are described below.

A variety of state diagrams ranging in size from 3 to 45 states were used to verify and collect statistical data on the algorithm. Fig. 4 shows the variation in runtime as the number of states in the diagram increased. A set of state diagrams was formed using the same number of states and varying their interconnecting paths. The time required to determine the regular expressions from this set of state diagrams was recorded. Changing the number of states, different sets of state diagrams with varying interconnecting paths were used. Later the times for each of separate sets of state diagrams were arranged in ascending order and graphed. After proper scaling, the results were found to be similiar. Fig. 5 shows this basic graph, where T is the least amount of time required to compute any regular expression of n states and the scale along the abscissa gives the proportion of regular expressions computed from the set of state diagrams of n states at some time xT, where $x > 1$. This increase in the amount of time required to compute the various regular expressions of the same number of states was considered to be an indicator of the complexity of that derived regular expression due to the algorithm. When the state diagram generated a regular expression of the form $R = E$ or EU^*, where E did not contain the universal sequence U^* which for the set $\{0,1\}$ is equal to $(0 + 1)^*$, the run time was almost minimal. In the case where the regular expression could have been written in the form $R = U^*E$, the derived regular expression incorporated the sequence E into the sequence U^*, increasing its complexity and the run time. Combining the information given in Fig. 4 and 5, we are able to determine the upper and lower boundaries for the run time which are given in Fig. 6

REFERENCES

[1] S. C. Kleene, "Representation of Events in Nerve Nets and Finite Automata," Automata Studies, C. E. Shannon and E. J. McCarthy, eds. Princeton University Press, Princeton, N.J., 1954, pp. 3-41.

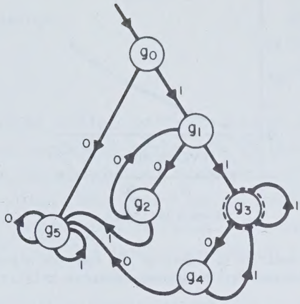
[2] I. M. Copi, C. L. Elgot, and J. B. Wright, "Realization of Events by Logical Nets," Jour. Assn. Computing Machinery, Vol. 5, pp. 181-196, April, 1958.

[3] R. McNaughton and H. Yamada, "Regular Expressions and State Graphs for Automata," IRE Trans. on Electronic Computers, Vol. EC-9, pp. 39-47, March, 1960.

[4] D. N. Arden, "Delayed Logic and Finite State Machines," Proc. 2nd Ann. Symp. on Switching Circuit Theory and Logical Design, Detroit, Mich., pp. 133-151, Oct.,1961.

[5] J. A. Brzozowski and E. J. McCluskey, "Signal Flow Graph Techniques for Sequential Circuit State Diagrams," IEEE Trans. on Electronic Computers, Vol. EC-12, pp. 67-76, April 1963.

[6] J. A. Brzozowski, "Derivatives of Regular Expressions," Jour. Assn. Computing Machinery, Vol. 11, No. 4, pp. 481-494, Oct., 1964.


g_0
1
0
g_1
0
0
1
g_2
g_3
1
0
g_5
1
0
1
0
g_4
1

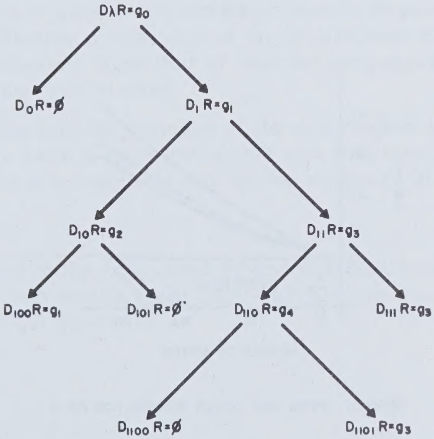
THE STATE DIAGRAM FOR THE ILLUSTRATIVE EXAMPLE

FIG. 1

	0	1	z
g_0	g_5	g_1	
g_1	g_2	g_3	
g_2	g_1	g_5	
g_3	g_4	g_3	x
g_4	g_5	g_3	
g_5	g_5	g_5	

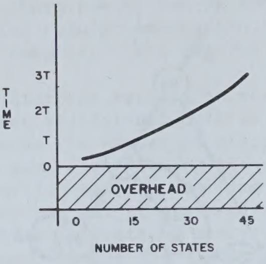
THE FLOW TABLE FOR FIG. 1

FIG. 2


$D_\lambda R = g_0$
$D_0 R = \emptyset$
$D_1 R = g_1$
$D_{10} R = g_2$
$D_{11} R = g_3$
$D_{100} R = g_1$
$D_{101} R = \emptyset$
$D_{110} R = g_4$
$D_{111} R = g_3$
$D_{1100} R = \emptyset$
$D_{1101} R = g_3$

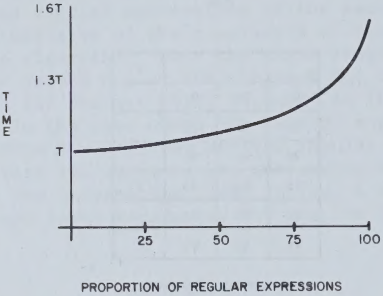
THE TREE FORMED FROM THE SET OF MINIMAL DERIVATIVES
OBTAINED IN STEP 1 FOR THE EXAMPLE

FIG. 3



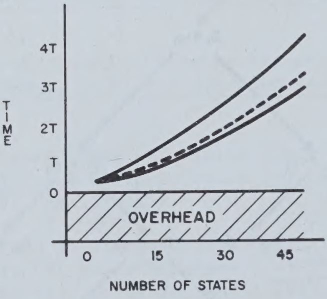
VARIATION IN COMPUTER RUN TIME FOR STATE DIAGRAMS WITH INCREASING NUMBERS OF STATES

FIG. 4



VARIATION IN COMPUTER RUN TIME FOR STATE DIAGRAMS WITH A CONSTANT NUMBER OF STATES BUT VARYING INTERCONNECTING PATHS

FIG. 5



DERIVED UPPER AND LOWER BOUNDS FOR FIG. 4

FIG. 6

ON AUTOMATA WITH FINITE REFERENCED PATTERNS

EDWARD J. KRALL
Dept. of Computer Science
University of Dayton
Dayton, Ohio

and

RAYMOND T. YEH
Electronic Research Ct
University of Texas
Austin, Texas

A class of finite automata called pattern-based automata is investigated in this paper. The acceptance criterion of these automata is based on matching patterns of the input against elements of a fixed finite pattern set. If the patterns mesh within the word being tested, the interesting class of locally acceptable languages[1] results. Formally, we define a k-acceptable language (kAL) as the language which is composed of words all of whose k-length subwords are in a finite pattern set. The specification of the pattern set completely determines the language. A locally acceptable language (LAL) is one which is kAL for some k.

This interesting subclass of the regular languages is a subclass of McNaughton and Papert's locally testable events[2] which are a generalization of the definite languages[3].

The class of kAL is also precisely the class of languages for which a word is in the language if and only if each i-length subword is in the language, for all i greater than or equal to k.

By allowing the pattern sets to contain words of differing lengths, one can obtain the pattern-based languages (PBL).[1] We have shown that except for a finite number of short words, each PBL is equal to a LAL. We also demonstrate finite procedure for deciding whether a language is LAL.

By allowing the automaton to check only every mth subword, for m greater than 1, and even by allowing the pattern to be broken at the end of the word and continued at the beginning, we obtain the skip-variant locally acceptable language (SVLAL) and the circular SVLAL[4], respectively. These rich classes are outside of McNaughton and Papert's class of Non-counting events. Closure properties of these different classes of languages (which are different from that of regular languages) and decision procedures have also been studied.

Thus by investigating the concept of the automation as a pattern-checking device, we have discovered some new and interesting subclasses of the regular languages, and thus add to the richness of that class of language.

This paper is partially supported by the National Science Foundation under grant GJ-786 and by the Joint Services Electronics Program under Research Grant AFSOR 69-1792.

I. PRELIMINARIES

A language L over al alphabet Σ is a subset of the free semigroup, Σ^*, generated by Σ. If α is a regular expression, then we use the symbol $\langle\alpha\rangle$ to denote the regular language represented by α. When no confusion arises, we do not distinguish between α and $\langle\alpha\rangle$ in general.

A finite automaton is ordered quintuple $(S, \Sigma, s_o, \delta, F)$ where S is a non-empty finite set of states, s_o is the start state, δ is the transition function: $\delta : S \times \Sigma \rightarrow S$, and F is a subset of S called the final state set.

If w is a word over some alphabet, then length of w, denoted by $\ell g(w)$, is the number of letters occured in w; if X is a set, then $|x|$ shall mean the number of elements in X. The complement of X is X^c. If w and v are words, then $w \leq v$ means that w is a subword of v, $w \leq_i v$ means it is an initial subword of v.

A language L is a k-testable language (kTL) (2) if for any two words w and v such that $L_k(w) = L_k(v)$, $R_k(w) = R_k(v)$, and $L_k(w) = L_k(v)$, then $w \in L$ if and only if $v \in L$; where L_k, R_k and L_k are functions mapping a word into its initial k-length subword, its final k-length subword, and the set of its internal k-length subwords. A language L is locally testable if and only if it is k-testable for some k, and will be denoted LTL. The class of languages denoted by LTO is the smallest class containing the LTL and closed under the boolean operations and concatenation.

A non-counting (2) language (denoted NC) is a regular language L such that for some n, and all words u, v, w over its alphabet, $uv^{n+i}w \in L$ if and only if $uv^n w \in L$, for all positive integers i.

A star-free (2) language (denoted SF) is one which is representable by a regular expression not containing the star operator.

It was shown by McNaughton and Paper (2) that LTO = NC = SF.

II. LOCALLY ACCEPTABLE LANGUAGES

An automaton which accepts the language, namely, LTO, SF and NC, studied by McNaughton and Papert (2) can be viewed as a pattern checking machine. Motivated from this point of view, we have obtained a new class of regular languages called the locally accepted languages which are recognized by very simple pattern checking automata. Relationships between LTO and this new class of language will also be mentioned.

Definition 2.1: A language $L \in \Sigma^*$ is called a k-acceptable language (kAL) with respect to a (finite) pattern set $P \subseteq \Sigma^k$ if and only if every k-length subword of elements of L is in the set P. We will usually denote L by the symbols L(P, k). L is called a locally-acceptable language (LAL) if it is a kAL for some k.

Theorem 2.1: A kAL is a regular language.

Proof: The regular expression for a kAL with respect to a pattern set P is $(\Sigma^* (P^c \cap \Sigma^k) \Sigma^*)^c$.||

Theorem 2.2: Except for a finite number of short words, every kAL is also (K + 1)AL.

Proof: Let L(P, k) be a kAL. Construct $Q = \Sigma^{k+1} \cap L(P,k)$. Clearly, L(P, k) and L(Q, k+1) differ by only a finite number of words of length less than k.||

Theorem 2.3: A language L is a kAL if and only if for every word $w \leq \Sigma^*$, $w \in L$ implies all of its subwords of length k or greater, are also in L and conversely.

Proof: Clearly if L is a kAL, then it is closed under k-length subword construction. Hence the condition of the theorem holds.

Conversely, suppose the condition of the theorem is satisfied. Let P be the set of all k-length subwords of words in L. Then it is easily seen that $L \subseteq L(P,k)$. On the other hand, if $w \in L(P,k)$ and $\ell g(w) = k$, then $w \in P$ and hence $w \in L$. Let us assume now that $w \in L(P,k)$ implies $w \in L$ for $k \leq \ell g(w) \leq m$. Let $w' \in L(P,k)$ such that $\ell g(w') = m + 1$. Then $w' = awb$, where $a, b \in \Sigma$ and $w \in \Sigma^{m-1}$. Since by hypothesis, $aw \in L$ and $bw \in L$, hence the condition of the theorem implies that $w' = awb \in L$. Thus $L(P,k) \subseteq L$ and the theorem holds.||

Theorem 2.4: The family of kAL is closed under intersection, ϵ-free homomorphism, word reversal, initial subwords, left and right derivative, and inverse homomorphism. It is not closed under union, complement, concatenation, star, arbitrary homomorphism, substitution, and intersection with regular languages.

Proof: Suppose L(P, k) and L(Q, k) are two given kAL's:

(a) Intersection: Assume $w = \sigma_1 \ldots \sigma_n$ $L(P,k) \cap L(Q,k)$. Then for each i, $\sigma_i \ldots \sigma_{i+k-1} \in P \cap Q$, which implies that $w = L(P \cap Q, k)$ and $L(P,k) \cap L(Q,k) \subseteq L(P \cap Q, k)$. By reversing the argument, we have $L(P,k) \cap L(Q,k) = L(P \cap Q, k)$.

(b) ϵ-free homomorphism: Let $w \in L(P,k)$. Then each k-length subword v of w is equal to some $p_r \in P$. Given some ϵ-free homomorphism $h: \Sigma \rightarrow \Gamma$, each k-length subword h(v) of H(w) is equal to some $h(p_r)$ of h(P). Thus $h(L(P,k)) = L(h(P),k)$.

(c) Word-reversal: Let $L^r = \{ w^r / w \in L \}$ denote the reversal of L. Clearly $L(P,k)^r = L(P^r, k)$.

(d) Initial subwords of length $\geq$ k: This is an immediate consequence of the definition. In fact, all subwords of length greater than or equal to k in L(P, k) are also in the language. As a consequence, kAL's are also closed under derivatives.

(e) Inverse homomorphism: If $h: \Gamma^* \rightarrow \Sigma^*$ is a homorphism, then $L(h^{-1}(P), k) = h^{-1}(L(P, k))$.

(f) Union: $L_1 = a^*b^* - \{\epsilon\} = L(\{ab, aa, bb\}, 2)$ and $L_2 = b^*a^* - \{\epsilon\} = L(\{ba, aa, bb\}, 2)$. However, $L_1 \cup L_2$ cannot be a 2AL for its pattern set must include $\{a, b\}^2$ which generates $\{a, b\}^* - \{\epsilon\}$.

(g) Complement: Let $L = L(\{aa\}, 2)$. Then $L^c = \{a\}^* - L = \{a, \epsilon\}$ cannot be a kAL for any k since, to include ϵ, k must be zero. But to include a, k must be 1.

(h) Concatenation and Star: Using the language L_1 in (f). Then $L_1L_1 = a^*b^*a^*b^* - \{\epsilon\}$ which is a proper subset of $\{a, b\}^*$. However, if L_1L_1 were a kAL, then its pattern set must be equal to $\{a, b\}^k$. However, $L(\{a, b\}^k, k) \neq L_1L_1$. Similar arguments apply to the star operator.

(i) Arbitrary homomorphism: The language $L = a^*cb^2b^*ca^*$ is a 3AL. However, if h is a homomorphism induced by the function $\binom{abc}{ab\epsilon}$, then $h(L) = a^*b^2b^*a^*$ cannot be a kAL, for any k.

(j) Substitution: Consider the 2AL, $L = aa^*b$ and the substitution S: $a \rightarrow a^*$, $b \rightarrow b^*a^*$, then $S(L) = a^*b^*a^*$, which is not a kAL, for any k.

(k) Intersection with regular languages: Consider the 2AL. $L_1 = a^* - \{\epsilon\}$ and the regular (but non-kAL) language $L_2 = \{a^2\}^*$. But since $L_1 \cap L_2 = L_2$, it is not kAL, for any K.

Since the arguments used in the previous theorem apply to LAL's, we have:

Corollary 2.5: The family of all LAL's have the same closure property as kAL's.

We remark that although LAL is not closed under concatenation, the following simple result, which we give here without proof, indicates that it is closed under a restricted form of concatenation.

Theorem 2.6: If $L = L(P, k)$ is a kAL over Σ, then $\Sigma L = L(\Sigma P, k+1)$.

Corollary 2.7: If $L = L(P, k)$, then $\Sigma^j L \Sigma^i = L(\Sigma^j P \Sigma^i, k+j+i)$, $i, j \geq 1$.

III. GENERAL PATTERN-BASED LANGUAGES

A generalization of the k-acceptable language is the general pattern-based language, in which the pattern set is not restricted to contain pattern words of one given length.

Definition 3.1: A language L is called a pattern-based language (PBL) with respect to a pattern set

$$P = \{p_1, \ldots, p_q\},\ p_i \in \Sigma^*,$$

if $P \subseteq L$ and for each arbitrary word $w = \sigma_1 \ldots \sigma_n \in \Sigma^*$, $w \in L$ if and only if for $i = 1, 2, \ldots, n - \min\limits_{1 \leq j \leq q} \lg(p_j) + 1$: there exists an r such that $\sigma_i \sigma_{i+1} \cdots \sigma_{i+m-1} = p_r$ with $m = \lg(p_r)$

We shall denote L by L(P).

We note from the previous definition that the existence of certain short patterns in a PBL forces it to include other short patterns of the same length. This is illustrated by the following example.

Example: Suppose P = {abaaa, ba, aa}.

Then abaaa $\in$ L(P) with the sequence of patterns: abaaa, ba, aa, aa, aa. Consider now a set Q = {abaaa, aa}. The first pattern word is still abaaa; but the second (ba, or baa, or baaa) is not in Q. Hence abaaa is both in

L(Q) and not in L(Q). This is a contradiction and hence Q cannot be a pattern set of any PBL. Because of the presence of the pattern as in P, the definition forces the pattern testing to be executed 4 times, for patterns beginning with a-, b-, a-, a-. If the pattern aa is eliminated, then we test only once.

Theorem 3.1: Corresponding to each PBL L, there exists an $\ell > 0$ such that for each $k \geq \ell$, there is a kAL L' which differs from L by only a finite number of (short) words.

Proof: Let L = L(P) for some pattern set P and let n be the length of the longest pattern in P. For $k \geq n$, let P_k be consisting of all the k-length subwords of L(P). Clearly, $L(P_k, k) = L(P) - (L(P) \cap \bigcup_{i=1}^{k-1} \Sigma^i)$.

Corollary 3.2: PBL's are regular.

Since the proof of the following theorem is similar to that of theorem 24, we will omit it here.

Theorem 3.3: The closure property of PBL is identical to that of LAL.

Lemma 3.4: Every kAL is a kTL for all $k \geq 0$.[1]

Proof: Let L be a given kAL, and let $w = \sigma_1 \sigma_2 \ldots \sigma_m$ and $w' = \sigma_1' \sigma_2' \ldots \sigma_n'$ be two words. Assume now that $R_k(w) = R_k(w')$, $L_k(w) = L_k(w')$ and $L_k w) = L_k(w')$. If $w \in L$, then clearly, $R_k(w)$, $L_k(w) \in P$ and $L_k(w) \leq P$ if and only if $w' \in L$. But this is precisely the definition of kTL. Hence, L is a kTL.||

Corollary 3.5: The class of LAL is a subclass of LTL.

Theorem 3.6: It is decidable whether a regular language is LAL.

Proof: Since it is decidable whether a regular language is an LTO we may, be corollary 35, assume that the language L in question is an LTO. Let N be the number of states of the reduced automation $A = (S, \Sigma, s_0, \delta, F)$ recognizing L. For each $k = 0, 1, 2, \ldots, N$, let P_k be the set of all k-length subwords of words in L. We claim that if

$$L(P_k, k) \cap (\bigcup_{i=1}^{k} \Sigma^i) = L \cap (\bigcup_{i=1}^{k} \Sigma^i) \tag{3}$$

then

$$L = L(P_k, k) \tag{4}$$

Otherwise, L is not kAL for this k.

Suppose, (3) holds, but (4) does not. Then there must be some word w, $\lg(w) \geq N$, such that either

(a) $w \in L$ and $w \notin L(P_k, k)$, or
(b) $w \in L(P_k, k)$ and $w \notin L$.

But (a) cannot be true by the construction of $L(P_k, k)$. Now, let ρ be right invariant equivalence relation on Σ^* such that

$$\rho = \{(w, v) \mid \delta(s_0, v) = \delta(s_0, v)\} .$$

If (b) holds, then there exists $w' \in \Sigma^*$ such that $(w, w') \in \rho$ and $\lg(w') \leq N$ and

$$w' \in L \cap L(P_k, k) \tag{5}$$

However, (b) and (5) cannot hold simultaneously. Since (5) is a consequence of (3), we have a contradiction. Hence the claim stands.||

---

[1] In 2 , McNaughton and Papert do not explicitly discuss short words i.e., words of length less than k + 1, in any kTL.

IV. SKIP-VARIANT LOCALLY-ACCEPTABLE LANGUAGES

By allowing the pattern testing of a word to occur in skips, the skip-variant LAL are constructed. If, in addition, the skips cause the testing to wrap around to the beginning of the word, the circular skip-variant LAL result.

Definition 4.1: A language L is called a *skip-variant kAL* with m-length skips with respect to a pattern set $P \leq \Sigma^k$ if for an arbitrary word $w = \sigma_1 \ldots \sigma_n \leftarrow \Sigma^*$, $w \in L$ if and only if

$$\sigma_i \ldots \sigma_{i+k-1} \in P \text{ for } i = 1,\ 1+m,\ 1+2m,\ \ldots\ 1+qm.$$

where q is the largest integer such that $1 + qm \leq n-k+m$. Such a language will be denoted L(P, k, m). A language is *skip-variant locally-acceptable* (SVLAL) if it is a SVkAL for some k.

A degenerate case exists when $k \leq m$: In this case, we have

$$L(P,k,m) = (P\Sigma^{m-k})^*.$$

Theorem 4.1: A SVkAL is a regular language.

Proof: Let L be a given SVkAL L(P,k, m). Then $L = \langle((\Sigma^m)^*(P^c \cap \Sigma^k)^*)^c\rangle$.||

Since the regular expression for a SVkAL with skip length greater than one contains a non-star-free subexpression, we have:

Corollary 4.2: The family of SVLAL is not contained in the family of NC languages.

Theorem 4.3: The closure properties for SVkAL and kAL are identical. The closure properties for SVLAL and LAL are identical, except that SVLAL is not closed under arbitrary homorphism.

Proof: The properties are easily verified by applying the some arguments as in Theorem 2.4. For arbitrary homomorphism however, not only would an ϵ-mapping destroy the pattern length in general, but also the word-length property of SVLAL: that only words of certain lengths are allowed in the language. Hence, the image language would have words of length, say, q, and also words of length q-1. Only a SVLAL with skip-length of 1 has this property; this implies the language is $h(\Sigma)^*$ for some h, and would not, in general, be equal to h(L(', k, m)). This fact is illustrated by considering the example where this pattern set P = {abab, acab, abac} and a homomorphism h such that 1(a) = a, $h(b) = \epsilon$ and h(c) = c. Clearly, $L(P, m, k) \neq h(L)$ for any m and k.

We will now show that SVLAL with wrap around testing is also regular. First, we need the following definition.

Definition 4.2: The *effective word multiple* of a word w is a function denoted by IBM such that for any non-negative integers k and m, EWM(w, k, m) is the least number r such that $\not\ell g(w)r = k (\text{mod } m)$.

Definition 4.3: A language L is a *circular SVkAL* with respect to a finite pattern set P and with skip length m if for an arbitrary word w

$$w \in L \text{ if and only if } w^{\mathrm{EWM}\ w,k,m)} \in L(P,k,m).$$

L will usually be denoted $1_c(P, k, m)$.

Note that it follows from the previous definition that some word w is in the circular SVkAL if and only if some multiple of w, depending on w, k, and m, is in the corresponding noncircular SVkAL.

Theorem 4.4: The circular SVLAL are regular.

Proof: By construction of a non-deterministic two-way automation which recognizes a given language $L_c(P, k, m)$.

Let $P = \{\sigma_{i1} \cdots \sigma_{ik}$, for $i = 1, 2, \ldots, q\}$. We define an automation A such that $A = (S, \Sigma, \delta, \{\bar{\sigma}_{11}, \bar{\sigma}_{21}, \ldots \bar{\sigma}_{q1}\}, F)$ where

$$S = \{\bar{\sigma}_{ij}, 1 \leq i \leq q, 1 \leq j \leq k\} \cup \{s_1, s_2, s_3, s_4\},$$

$$\delta: \quad S \times \Sigma \rightarrow S \times \{r, o, \ell\},$$

$$F = \{s_2\}.$$

We shall make two simplifying assumptions:

1) The read-head can move more than one square at a time; for example, $(s, 2r)$ would mean change state to s, and move 2 squares to the right.
2) The left and right ends are recognized, and any movement off those ends is automatically continued at the other end and shall be the end marker.

Clearly, these extensions can be accounted for by adding word states. (See Figures 4.2 and 4.3.) With these remarks in mind, we define the transition function.

$$\delta(\bar{\sigma}_{ij}, \sigma_{ij}) = (\bar{\sigma}_{ij+1}, r), \quad 1 \leq i \leq q, \quad 1 \leq j \leq k-1.$$

$$\delta(\bar{\sigma}_{ik}, \sigma_{ik}) = (s_1, r), \quad 1 \leq i \leq q.$$

$$\delta(s_1, \text{¢}) = (s_2, \text{halt}).$$

$$\delta(s_1, \sigma) = (s_3, (k-m+1)\ell), \quad 1 \leq i \leq q. \tag{8}$$

$$\delta(s_3, \text{¢}) = (s_4, \text{halt}). \tag{9}$$

$$\delta(s_3, \ ) = \{(\bar{\sigma}_{i1}, 0) \mid 1 \leq i \leq q\}. \tag{10}$$

After a successful pattern match, (8) forces the read-head to move to the beginning of the next subword to be tested. At this point, (9) causes the tape to be examined for a ¢, a condition that will occur if the EWM function has no value for this word.

Assume $w = \sigma_1 \cdots \sigma_n \in L_c(P, k, m)$.

Then the subwords tested will be

$$\sigma_1 \cdots \sigma_k, \ \sigma_{1+m} \cdots \sigma_{k+m}, \ldots, \sigma_{n-k+1} \cdots \sigma_n.$$

Thus at the k, 2k, ... time unit, the automation is in state s_1. At the last testing, we will have the transition $\delta(s_1, \text{¢})$ at which point the word is accepted. Otherwise, transitions (8) and (10) will be effected. Transition (9) can never occur, for it would imply that $\sigma_{1+ax} = \text{¢}$ for some x; but this can only happen for words of prohibited length, that is, one in which the EWM has no value.

If on the other hand, the word $w \notin L_c(P, K, M)$ then either

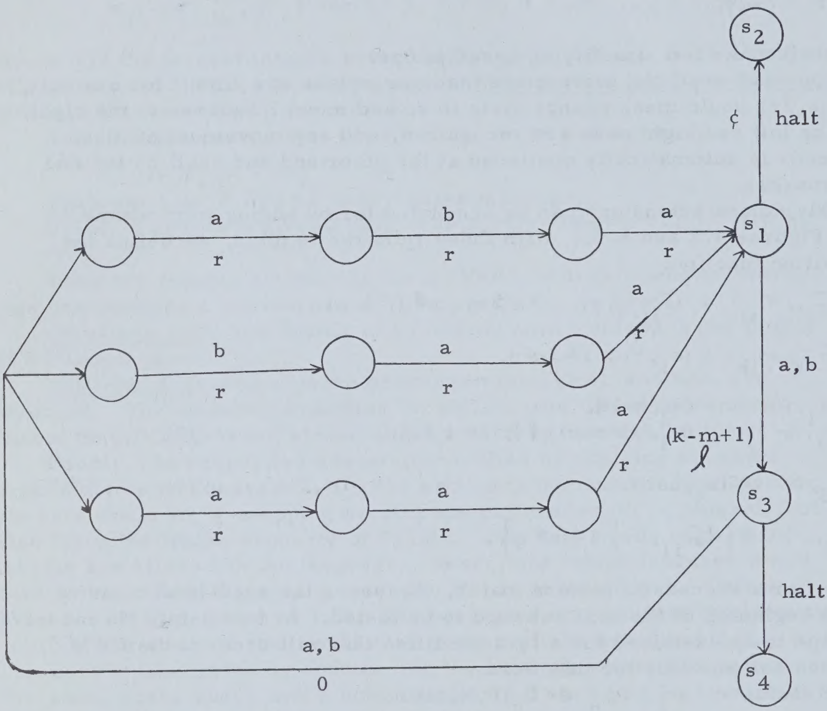
1) it is of a prohibited length, in which case transition (9) will trap it, or
2) some subword will not match any pattern, in which case, the automation will "hang" before reaching

state s_1 for that test.

An example illustrating the previous theorem is given below.

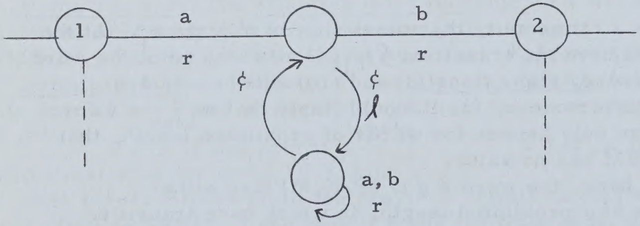
Example: Let $L = L_{¢}(P, 3, 2)$, where $P = \{aba, baa, aaa\}$. Then the state-diagram of the (non-deterministic) two-way automation recognizing L is given in figure 4. 1.

The hidden states (caused by the simplified assumptions in the theorem) which effect the wrap-around feature are shown in figures 4. 2 and 4. 3 for illustrative purposes.



s2
¢ halt
a
r
b
r
a
r
s1
a
r
b
r
a
r
a, b
a
(k-m+1)
ℓ
r
a
r
a
r
s3
halt
a, b
0
s4

Figure 6. 1 - State-diagram of a Two-Way Automation Recognizing a Circular SVkAL.



1
a
r
b
r
2
¢
¢
ℓ
a, b
r

Figure 6. 2 - Hidden States Between States 1 and 2 in Figure 6. 2.

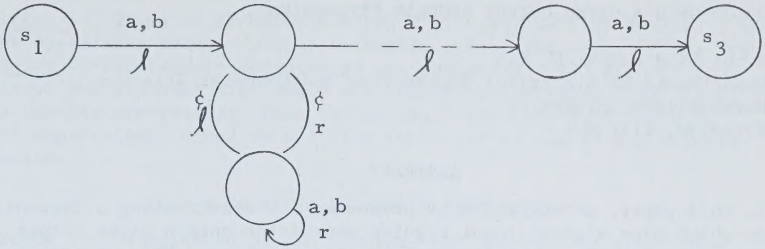


Figure 6.3 - Hidden States Between s_1 and s_3 in Figure 6.1.

Finally, we know that it is decidable to determine if a given language is LAL, and of course, LTL and NC. The existence of a decision procedure for SVLAL has eluded us, but we suspect that the question is decidable.

	kAL	LAL	PBL	SVkAL	SVLAL
Union	N	N	N	N	N
Intersection	Y	Y	Y	Y	Y
Complementation	N	N	N	N	N
Concatenation	N	N	N	N	N
Star	N	N	N	N	N
Homomorphism	N	Y	Y	N	N
-free Homomorphism	Y	Y	Y	Y	Y
Inverse Homomorphism	Y	Y	Y	Y	Y
Substitution	N	N	N	N	N
Reversal	Y	Y	Y	Y	Y
Derivative	Y	Y	Y	Y	Y
Initial subwords	N	N	N	N	N
Initial subwords of length greater than or equal to k.	Y	Y	Y	Y	Y

Table 9.1: Closure Properties of Various Languages

REFERENCES

1. Krall, Edward J. A Classification of the Regular Language. PhD Dissertation, University Microfilms (1971)

2. McNaughton, R., and Papert, S. Computer-free Automata. Manuscript, to be published in 1971.

3. Perles, M., Rabin, M. O. and Shamir, E. "The Theory of Definite Automata", IEEE Trans. Electronics Computers EC-12, 233-243, 1963.

CONSTRUCTION OF A SEQUENTIAL MACHINE MAPPING A GIVEN INPUT REGULAR EXPRESSION ONTO A GIVEN OUTPUT REGULAR EXPRESSION

R. O. Lind and S. S. Yau
Departments of Electrical Engineering and Computer Sciences
Northwestern University
Evanston, Illinois

ABSTRACT

In this paper, an algorithm is presented for constructing a sequential machine which maps a given input regular expression onto a given output regular expression if such a sequential machine exists. This algorithm is based on the properties of transition matrices and regular expression.

I. INTRODUCTION AND DEFINITIONS

The relationship between regular expressions and sequential machines has been studied extensively, since Kleene's introduction of regular expressions [1]. Many authors have investigated the conversion of representatives of sequential machines between regular expressions and other well known forms, such as state diagrams or flow tables [2-5]. Concerning the problem of finding the image of regular expressions for mapping through a sequential machine, two methods have been presented. One method was presented by Yau [6] using the regular input expression and the state graph of the sequential machine to construct the output graph. The other method using transition matrices was presented by Kameda [7]. Ginsburg and Hibbard [8] showed that there exists a complete sequential machine (csm) which maps a regular expression into a regular expression. Furthermore, they showed that it is recursive solvable to determine whether or not there exists a csm which maps a regular expression onto a regular expression. However, there is no reasonable algorithm to construct such a csm if there exists one. In this paper, we shall present an algorithm for constructing a sequential machine which maps a given input regular expression onto a given regular output expression if such a sequential machine exists. This algorithm is based on the properties of transition matrices and regular expressions.

Before the main topic of this paper is discussed, several definitions must be given.

Definition 1. A Mealy type sequential machine M is a quintuple

$$M = (S,I,O,f,g)$$

where $S = \{s_1,s_2,\ldots,s_u\}$ is a finite state set, $I = \{i_1,i_2,\ldots,i_v\}$ is the finite input alphabet, $O = \{o_1,o_2,\ldots,o_r\}$ is the finite output alphabet, $f{:}S \times I \rightarrow S$ is the next state function, and $g{:}S \times I \rightarrow O$ is the output function.

In this paper, a reference to a sequential machine will mean a Mealy type sequential machine.

For the discussion of the algorithm, the definition for an acceptor, usually called an automaton, will be needed.

Definition 2. An acceptor is a 5-tuple

$$M = (S,I,f,p_1,F)$$

where S is a finite nonempty set of states, I is a finite nonempty set of inputs, f is a mapping of $S \times I$ into S, p_1 is the starting state, and F is the set of final states.

Since the paper is concerned with regular expressions, the definition of regular expressions is required.

Definition 3. Given an alphabet $A = \{a_1, a_2, \ldots, a_l\}$ and the symbols λ and $\emptyset$; and the operations union, "$\cup$"; concatenation, "$\cdot$"; and star, "$*$"; a regular expression is recursively definied as the following: 1) Each of the symbols of the alphabet A, λ, and $\emptyset$ are regular expressions. 2) If S_1 and S_2 are regular expressions, then so are $(S_1 \cup S_2)$, $(S_1 \cdot S_2)$, and $S_1{*}$. 3) All expressions formed by a finite number of 1) and 2) are regular expressions.

II. T-MATRICES AND THEIR PROPERTIES

Since in the solution of the problem, the approach of transition matrices [7] will be used, then the definition of T-matrices and their properties will be discussed.

Definition 4. The transition-output matrix of a sequential machine M is an n x n matrix of the form

$$\begin{bmatrix} t_{00} & t_{01} & & t_{0\ n-1} \\ t_{10} & t_{11} & & t_{1\ n-1} \\ \vdots & \vdots & & \vdots \\ t_{n-1\ 0} & t_{n-1\ 1} & \cdots & t_{n-1\ n-1} \end{bmatrix} \qquad (1)$$

where n is the number of states of M, t_{ij} $(i,j = 0,1,2,\ldots,n-1)$ is the union of the input-output pair of the form (i,o) if and only if the input i takes M from s_i to s_j producing output o and if there is no input i taking M from s_i to s_j, then $t_{ij} = (\emptyset,\emptyset)$ where $\emptyset$ is the empty set.

Definition 5. A T-matrix of a sequential machine M is of the form of (1) where t_{ij} is a regular expression over the input (or output) alphabet such that all the sequences denoted by the regular expression takes M (or, all the sequences denoted by the regular expression are produced when M goes) from s_i to s_j.

Definition 6. Let $i \in I$, the input alphabet. The T-matrix ${}^iT$ of a machine M is the T-matrix obtained from the transition-output matrix of M by retaining the output symbols of its input-output pair if and only if their corresponding input symbol is i and all other entries are $\emptyset$, the empty set.

An example of the transition-output matrix for the sequential machine N is given in Fig. 1b. An example of the two T-matrices for machine N is given in Fig. 1c.

	0	1
s_0	s_2,0	s_1,1
s_1	s_1,1	s_0,1
s_2	s_2,0	s_1,1

(a)

$$\begin{bmatrix} (\emptyset,\emptyset) & (1,1) & (0,0) \\ (1,1) & (0,1) & (\emptyset,\emptyset) \\ (\emptyset,\emptyset) & (1,1) & (0,0) \end{bmatrix}$$

(b)

$$\underset{{}^0T}{\begin{bmatrix} \emptyset & \emptyset & 0 \\ \emptyset & 1 & \emptyset \\ \emptyset & \emptyset & 0 \end{bmatrix}} \qquad \underset{{}^1T}{\begin{bmatrix} \emptyset & 1 & \emptyset \\ 1 & \emptyset & \emptyset \\ \emptyset & 1 & \emptyset \end{bmatrix}}$$

(c)

Fig. 1. (a) Machine N, (b) transition-output matrix of machine N, and (c) ${}^iT$ for machine N.

Now that the definitions have been given, the operations on T-matrices will be discussed.

Definition 7. The union of two T-matrices is a T-matrix whose (i,j) element is the union of the (i,j) element of the two T-matrices and this is denoted by $\cup$. The multiplication of two T-matrices is a T-matrix as is defined in matrix calculus with multiplication and addition of elements interpreted as those of concatenation and union. This is denoted by $\bullet$. The closure of a T-matrix T is denoted by the expression

$$T^* = E \cup T \cup T^2 \cup T^3 \cup \ldots \tag{2}$$

where E is the unit T-matrix whose diagonal elements are the null sequences λ and off-diagonal elements are $\emptyset$. This is denoted by T^*.

With the definition of the operations, the only difficult part is how to find the closure of T-matrix. The closure can be found by using Lemma 1 and (5).

Lemma 1 (Arden): Let R and S be regular expressions such that $\lambda \notin R$. Then the equation

$$X = XR \cup S \tag{3}$$

has the unique solution

$$X = SR^* \tag{4}$$

Lemma 2 (Kameda): Let X_{ij} be the (i,j) element of T^*, the closure of the T-matrix T. Then there are the following relations for each i:

$$\begin{aligned} X_{ii} &= \bigcup_{k=0}^{n-1} X_{ik} a_{ki} \cup \lambda \\ X_{ij} &= \bigcup_{k=0}^{n-1} X_{ik} a_{kj} \qquad \text{for } i \neq j \end{aligned} \tag{5}$$

where n is the number of states of machine M and a_{kj} is the (k,j) element of the T-matrix T for machine M.

Now that the definition of transition matrices have been given, the relationships between T-matrices and regular expressions will be shown.

Definition 8. A T-matrix is said to correspond to a regular event R over the input (or output) alphabet, if and only if the (i,j) element of the T-matrix denotes all the output (or input) sequences, each of which is produced when M goes (or, each of which takes M) from s_i to s_j, produced from an input (or, producing an output) sequences in the event R. We define the following correspondence.

$$\lambda \longleftarrow E = \begin{bmatrix} \lambda & \cdots & \emptyset \\ & \lambda & \\ & & \ddots \\ \emptyset & & \lambda \end{bmatrix} \qquad \emptyset \longleftarrow \Phi = \begin{bmatrix} \emptyset & \cdots & \emptyset \\ \vdots & & \vdots \\ \emptyset & & \emptyset \end{bmatrix}$$

Theorem 1 (Kameda): Let R be a regular expression over the input alphabet. The T-matrix corresponding to $|R|$ is equal to $R[{}^iT]$, which is obtained by replacing the input symbol i by ${}^iT$ for each i, λ by E, and $\emptyset$ by Φ.

III. NUMBER OF STATES

Looking at the definition of transition matrices, one sees that the size of a transition matrix depends upon the number of states of the sequential machine. Therefore, when transition matrices are used, the number of states of the sequential machine must be known. Since, the input and output regular expressions are all that is given, then a number of states in a sequential machine must be determined from this information.

Before showing how to determine an initial number for the number of states from the input and output regular expressions, we must define a unit subsequence.

Definition 9. The unit subsequence is the concatenation of a number of input (or output) symbols without involving any other operations in an input (or output) regular expression.

As an example of a unit subsequence expression, consider the regular expression below.

$$R = (0 \cup 10)^*1$$

From this regular expression and the definition of a unit subsequence, we obtain that there are three unit subsequences which are

0, 10, and 1

Now that the definition of the unit subsequence has been defined, the procedure for determining an initial number for the number of states of a sequential machine from the given input and output regular expressions can be given. The procedure is:

i) Determine the number of unit subsequences in the input regular expression and in the output regular expression denoted by U_i and U_o respectively.

ii) Divide the number of unit subsequences in the output regular expression, U_o, by the number of unit subsequences in the input regular expression, U_i.

iii) Set n, the number of states, equal to an integer near the number obtained in step ii). The value of n must be greater than one.

The above procedure gives an initial value for the number of states to be used in the algorithm to be given with which to begin the determination of a sequential machine from the input and output regular expressions. It is known that there are usually several regular expressions which represent the same regular set of sequences. Because of this fact, it is possible to obtain several different initial values for the number of states by the above procedure. It is also known that it is sometimes impossible or at most very difficult to transfer from one regular expression to another even if it is known that they are equivalent. For this reason and that we are looking for an initial value for the number of states, the above procedure is proper to give a starting point in the algorithm.

Another problem which must be considered is what to do if no sequential machine is found at the end of the application of the algorithm. What is to be done, is to increase the number of states by one and then reapply part of the algorithm. We must also consider how much to increase the number of states, in other words, what is the upper limit on the number of states. The upper limit on the number of states is equivalent to the product of the number of states of the acceptor which is determined from the input regular expression and the number of states of the acceptor which is determined from the output regular expression. This upper limit is denoted by mk where m is the number of states of the acceptor determined from the input regular expression and k is the number of states of the acceptor determined from the output regular expression. Now that the upper limit on the number of states has been stated, we must prove that it is the upper limit.

Theorem 2: If there exists a sequential machine which maps the input regular expression, R_i, onto the output regular expression, R_o, then such a sequential machine with mk states can always be found, where m and k are the number of states of the acceptors accepting R_i and R_o respectively.

Proof: Given the input and output regular expressions, R_i and R_o. Let $M_1 = (S_1,I,f_1,p_1,F_1)$ be the acceptor which is determined from the input regular expression with m states such that

$$R_i = \{i_1 i_2 \ldots i_r \mid f_1(p_1, i_1 i_2 \ldots i_r) \text{ is in } F_1\}$$

and let $M_2 = (S_2,0,f_2,q_1,F_2)$ be the acceptor which is determined from the output regular expression with k states such that

$$R_o = \{o_1 o_2 \ldots o_j \mid f_2(q_1, o_1 o_2 \ldots o_j) \text{ is in } F_2\}.$$

Let μ be the mapping of the input regular expression, R_i, onto the output regular expression, R_o. Let $s = (p,q)$ where p is in S_1 and q is in S_2 for which there exists an input sequence J in R_i such that $f_1(p_1,J) = p$ and $f_2(q_1,\mu(J)) = q$. Let S be the set of all possible pairs $(p,q) = s$. Now then for each s in S, let $L(p,q)$ denote an input sequence in R_i such that $f_1(p_1,L(p,q)) = p$ and $f_2(q_1,\mu(L(p,q))) = q$. For each s in S and each i in I, we let

i) $f(s,i) = (f_1(p,i),\ f_2(q,\mu_{L(p,q)}(i)))$ and

ii) $g(s,i) = \mu_{L(p,q)}(i)$

where $\mu_{L(p,q)}(i)$ is the last output in the output sequence of $\mu(L(p,q)i)$. The resulting sequential machine $M = (S,I,0,f,g,(p_1,q_1))$ is the sequential machine which maps the input regular expression onto the output regular expression. By construction, we get that the sequential machine has mk states. Therefore if there exists a mapping, μ, of the input regular expression, R_i, onto the output regular expression, R_o, there is a sequential machine with mk states. Thus the proof of the theorem.

This theorem proves that if there exists a sequential machine which maps the input regular expression onto the output regular expression, then the sequential machine has at most mk states. Now we must prove that this sequential machine can always be found using the algorithm. But first we must give the algorithm.

IV. ALGORITHM

The algorithm for determining a sequential machine when the input and output regular expressions are known is:

i) Determine n, an initial value for the number of states, as previously described using the number of unit subsequences in the input and output regular expressions.

ii) With the value for n, write the T-matrices, ${}^{i}T$, for each input i with variables for each entry in ${}^{i}T$.

iii) Consider the input regular expression and calculate $R[{}^{i}T]$ as has been described.

iv) Take the union of the terms of the first row of the final T-matrix, $R[{}^{i}T]$.

v) Compare the result of Step iv) with the given output regular expression and determine what each variable of the T-matrices are with the following limitations.

a) There are at most n terms in any row of all T-matrices which have value, not Ø. All other entries in the row of all T-matrices are Ø.

b) There can be no more than one entry for each row of each input T-matrix. The other entries of the row are Ø.

vi) From the T-matrices, ${}^{i}T$, the transition-output matrix can be determined. From the transition-output matrix, the state diagram of a sequential machine can be found.

vii) If no sequential machine can be found, the number of states, n, is increased by one and then go back to Step ii). If no sequential machine has been found with n at the upper limit, then it is to be considered that no sequential machine exists for the given input and output regular expressions.

The assumption made in this algorithm is that the input and output alphabets are known. It also should be noted that a sequential machine which is determined using the given algorithm, may not be unique. That is for the one set of given input and output regular expressions, there may be several different sequential machines obtained.

Now that the algorithm has been given, it must be proven that if there exists a sequential machine which maps the input regular expression onto the output regular expression, then such a sequential machine can always be found using the algorithm.

Theorem 3. Let m and k be the number of states of the acceptors accepting the input and output regular expressions. If there exists a sequential machine with mk states which maps the input regular expression onto the output regular expression, then a sequential machine with no more than mk states can be obtained using the algorithm.

Proof: Assume that there exists a sequential machine with mk states which maps the input regular expression onto the output regular expression. Since the sequential machine exists, then the T-matrices, ${}^{i}T$, for each input i can be determined. Also using the ${}^{i}T$'s, the output regular expression will be obtained from the calculated $R[{}^{i}T]$ as in [7].

Now calculate $R[{}^{i}T]$ with a variable in each entry of each input i T-matrix, ${}^{i}T$, as is done in the algorithm. Now take the union of the terms in the first row of the final T-matrix, $R[{}^{i}T]$, and will call this the result. Since the set of ${}^{i}T$'s exists for a sequential machine, therefore when the output regular expression is compared with the result, we will obtain a value, either an output symbol or Ø, for each variable in each input T-matrix. Thus a sequential machine is obtained with mk states using the algorithm. Therefore proof of the theorem.

Therefore from Theorems 2 and 3, we see that if no sequential machine was found for the number of states equal to the upper limit on the number of states, then it is to be considered that there is no sequential machine which maps the input regular expression onto the output regular expression.

V. EXAMPLES

Now that the algorithm has been given, several examples using the algorithm will be worked to show how the problem is solved.

Example 1: Given the input and output alphabet, I = 0 = {0,1}, and the input and output regular expressions.

$$R_i = (0 \cup 10)^*1$$

$$R_o = 0(00)^*0 \cup 1 \cup (11 \cup 0(00)^*0)(1 \cup 00)^*0$$

Step 1: First determine an initial value for n, the number of states by finding U_i and U_o. From this information, a value of 3 is obtained for n.

Step 2: Now construct the T-matrices for each input.

$$ {}^{0}T = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \qquad {}^{1}T = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix} $$

Step 3: Considering the input regular expression, we calculate

$$ R[{}^{i}T] = ({}^{0}T \cup {}^{1}T\,{}^{0}T)^{*}\,{}^{1}T $$

$$ {}^{0}T \cup {}^{1}T\,{}^{0}T = \begin{bmatrix} A_{11}\cup B_{11}A_{11}\cup B_{12}A_{21}\cup B_{13}A_{31} & A_{12}\cup B_{11}A_{12}\cup B_{12}A_{22}\cup B_{13}A_{32} & A_{13}\cup B_{11}A_{13}\cup B_{12}A_{23}\cup B_{13}A_{33} \\ A_{21}\cup B_{21}A_{11}\cup B_{22}A_{21}\cup B_{23}A_{31} & A_{22}\cup B_{21}A_{12}\cup B_{22}A_{22}\cup B_{23}A_{32} & A_{23}\cup B_{21}A_{13}\cup B_{22}A_{23}\cup B_{23}A_{33} \\ A_{31}\cup B_{31}A_{11}\cup B_{32}A_{21}\cup B_{33}A_{31} & A_{32}\cup B_{31}A_{12}\cup B_{32}A_{22}\cup B_{33}A_{32} & A_{33}\cup B_{31}A_{13}\cup B_{32}A_{23}\cup B_{33}A_{33} \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{bmatrix} $$

Using Lemma 2 for i = 0 and Lemma 1 with the proper substitution, we obtain

$$ \begin{aligned} X_{02} &= X_{00}D_{13}D_{33}^{*} \cup X_{01}D_{23}D_{33}^{*} \\ X_{01} &= X_{00}(D_{12} \cup D_{13}D_{33}^{*}D_{32})(D_{22} \cup D_{23}D_{33}^{*}D_{32})^{*} \\ X_{00} &= (D_{11} \cup D_{13}D_{33}^{*}D_{31} \cup (D_{12} \cup D_{13}D_{33}^{*}D_{32})(D_{22} \cup \\ &\qquad D_{23}D_{33}^{*}D_{32})^{*}(D_{21} \cup D_{23}D_{33}^{*}D_{31}))^{*} \end{aligned} $$

Since in the final result we are only concerned with the first row, we will only have to calculate the first row of the closure. Therefore the final T-matrix is

$$ \begin{bmatrix} X_{00}B_{11}\cup X_{01}B_{21}\cup X_{02}B_{31} & X_{00}B_{21}\cup X_{01}B_{22}\cup X_{02}B_{32} & X_{00}B_{13}\cup X_{01}B_{23}\cup X_{02}B_{33} \\ ? & ? & ? \\ ? & ? & ? \end{bmatrix} $$

Step 4: Now take the union of the first row to get the result which is finally after substitution and combining terms the following.

$$ \begin{aligned} \text{Result} = X_{00}(&B_{11}\cup B_{12}\cup B_{13}\cup D_{13}D_{33}^{*}(B_{31}\cup B_{32}\cup B_{33})\cup(D_{12}\cup D_{13}D_{33}^{*}D_{32})(D_{22}\cup \\ &D_{23}D_{33}^{*}D_{32})^{*}(B_{21}\cup B_{22}\cup B_{23}\cup D_{23}D_{33}^{*}(B_{31}\cup B_{32}\cup B_{33}))) \end{aligned} $$

Step 5: Now compare the above result with the regular output expression and we immediately see that

$$X_{00} = \lambda$$

and the remaining part of the result is equivalent to the output regular expression. By continuing the comparison, we obtain that

$$D_{33} = 00$$

$$D_{13} = 0$$

$$B_{31} \cup B_{32} \cup B_{33} = 0$$

$$B_{11} \cup B_{12} \cup B_{13} = 1$$

Substituting what we know, we get that

$$\text{Result} = 1 \cup 0(00)^*(0 \cup (D_{12} \cup 0(00)^* D_{32})(D_{22} \cup D_{23}(00)^* D_{32})^*(B_{21} \cup B_{22} \cup B_{23} \cup D_{23}(00)^* 0)$$

From the above expression, we obtain that

$$D_{12} = 11$$

$$D_{32} = 0$$

$$D_{22} = 1 \cup 00$$

$$D_{23} = \emptyset$$

$$B_{21} \cup B_{22} \cup B_{23} = 0$$

Since $X_{00} = \lambda = (\emptyset)^*$, we get that

$$\emptyset = D_{11} \cup 0(00)^* D_{31} \cup (11 \cup 0(00)^* 0)(1 \cup 00)^* D_{21}$$

and hence

$$D_{11} = D_{21} = D_{23} = D_{31} = \emptyset$$

Continuing on with the above information, we are able to determine a value, either an output symbol or $\emptyset$, for each of the variables in the T-matrices. With this information, we can construct the T-matrices which are

$${}^{0}T = \begin{bmatrix} \emptyset & \emptyset & 0 \\ \emptyset & 1 & \emptyset \\ \emptyset & 0 & \emptyset \end{bmatrix} \qquad {}^{1}T = \begin{bmatrix} \emptyset & 1 & \emptyset \\ \emptyset & \emptyset & 0 \\ 0 & \emptyset & \emptyset \end{bmatrix}$$

Step 6: From the above T-matrices, we get the transition-output matrix

$$\begin{bmatrix} (\emptyset,\emptyset) & (1,1) & (0,0) \\ (\emptyset,\emptyset) & (0,1) & (1,0) \\ (1,0) & (0,0) & (\emptyset,\emptyset) \end{bmatrix}$$

From the transition-output matrix, we can get the state diagram for the sequential machine which is shown below:

	0	1
s_0	s_2,0	s_1,1
s_1	s_1,1	s_2,0
s_2	s_1,0	s_0,0

Example 2: Given the input and output alphabet of {0,1} and the input and output regular expressions

$$R_i = (1 \cup 0)^*1$$

$$R_o = (0 \cup 0(1 \cup 0))^*(0 \cup 00)$$

First we find that an initial value for n is 2. Then we construct the T-matrices and calculate

$$R[{}^iT] = ({}^1T \cup {}^0T)^*\,{}^1T$$

From the final T-matrix $R[{}^iT]$, we take the union of the first row which gives us as a result.

$$\text{Result} = X_{00}(B_{11} \cup B_{12} \cup (A_{12} \cup B_{12})(A_{22} \cup B_{22})^*(B_{21} \cup B_{22}))$$

where

$$X_{00} = (A_{11} \cup B_{11} \cup (A_{12} \cup B_{12})(A_{22} \cup B_{22})^*(A_{21} \cup B_{21}))^*$$

Now comparing the result with the output regular expression and proceeding to determine the value for each variable in the T-matrices, it is found that two different sets of T-matrices are realized. They are shown below.

$$ {}^0T = \begin{bmatrix} 0 & \emptyset \\ 1 & \emptyset \end{bmatrix} \quad {}^1T = \begin{bmatrix} \emptyset & 0 \\ 0 & \emptyset \end{bmatrix} \qquad {}^0T = \begin{bmatrix} \emptyset & 0 \\ 1 & \emptyset \end{bmatrix} \quad {}^1T = \begin{bmatrix} 0 & \emptyset \\ 0 & \emptyset \end{bmatrix} $$

From the two sets of T-matrices, two transition-output matrices are obtained.

$$ \begin{bmatrix} (0,0) & (1,0) \\ (0,1) \cup (1,0) & (\emptyset,\emptyset) \end{bmatrix} \qquad \begin{bmatrix} (1,0) & (0,0) \\ (0,1) \cup (1,0) & (\emptyset,\emptyset) \end{bmatrix} $$

From the transition-output matrices, we obtain the two sequential machines which performs the desired mapping.

	0	1
s_0	s_0,0	s_1,0
s_1	s_0,1	s_0,0

	0	1
s_0	s_1,0	s_0,0
s_1	s_0,1	s_0,0

Example 3: Given the input and output alphabets of {0,1} and the input and output regular expressions

$$R_i = (1 \cup 0)^*1$$

$$R_o = (0 \cup 01^*0)^*0$$

First we determine that an initial value for n is 2. Proceeding through the algorithm, it is determined that no sequential machine exists with number of states at 2 which maps the input regular expression onto the output regular expression. We now increase the number of states by one so that now n = 3. We now proceed through the algorithm with n = 3. Again we find that no sequential machine was found to map the input regular expression onto the output regular expression. So again we increase the number of states by one and proceed through the algorithm with n now equal to 4. But again we find that there exists no sequential machine with 4 states which performs the desired mapping. Now since no sequential machine has been found for several values for the number of states, the upper limit, mk, should be determined. Determining the acceptors from the input regular expression and the output regular expression, it was found that both acceptors have 3 states, thus m = k = 3. Therefore the upper limit on the number of states is equal to 9.

Now applying the algorithm for the value of the number of states at 5, 6,7,8, and 9 results again in the condition that there exists no sequential machine for the given input and output regular expressions for the various number of states. Since the upper limit is equal to 9 and it was found that no sequential machine with the number of states equal to 9 exists for the given input and output regular expressions, it can therefore be said that no sequential machine exists which maps the input regular expression onto the output regular expression.

VI. CONCLUSION

In this paper, we have used the concept of transition matrices to establish algorithms for constructing a csm which maps a given input regular expression onto a given output regular expression. It was proven that if no sequential machine was determined for the upper limit on the number of states, then there exists no sequential machine which maps the input regular expression onto the output regular expression. Therefore, a sequential machine can always be found by the algorithm if one does exist.

In this paper, the sequential machines under consideration are of the Mealy type. But, the algorithm presented in this paper can equally apply to the use of the Moore type sequential machine with just a few minor modifications.

Since we are working with the input and output regular expressions, there may be more than one sequential machine which maps the input regular expression onto the output regular expression. Thus, the results of the algorithm may not be unique as was shown in Example 2. In this example, we obtained two different sequential machines which map the same input regular expression onto the same output regular expression.

REFERENCES

1. S. C. Kleene, "Representation of events in nerve nets and finite automata", Automata Studies, Annals of Math. Studies, No. 34, Princeton University Press, 1956, pp. 3-41.
2. R. McNaughton and H. Yamada, "Regular expressions and state graphs for automata", IRE Trans. on Electronic Computers, Vol. EC-9, No. 1, pp. 39-47, March 1960.
3. D. N. Arden, "Delayed-logic and finite-state machines", Proc. 2nd Ann. Symp. on Switching Circuit Theory and Logical Design, Chicago, Ill., pp. 131-151, September 1961.
4. G. Ott and N. H. Feinstein, "Design of sequential machines from their regular expressions", Jour. Assoc. Computing Machinery, Vol. 8, No. 4, pp. 585-600, October 1961.
5. J. A. Brozozowski and E. J. McCluskey, "Signal flow graph techniques

for sequential circuit state diagrams", IEEE Trans. on Elec. Computers, Vol. EC-12, No. 2, pp. 67-77, April 1963.

6. S. S. Yau, "Generation of an output regular expression of a sequential machine with a specified input regular expression", Proceedings of the 4th Annual Allerton Conference on Circuits and System Theory, 1966, University of Illinois, Urbana, Illinois, pp. 343-351.
7. T. Kameda, "Generalized transition matrix of a sequential machine and its application", Information and Control, Vol. 12, No. 3, pp. 259-275, March 1968.
8. S. Ginsburg and T. N. Hibbard, "Solvability of machine mappings of regular sets to regular sets", Jour. Assoc. Computing Machinery, Vol. 11, No. 3, pp. 302-312, July 1964.
9. S. Seshu, R. E. Miller and G. Metze, "Transition matrices of sequential machines", IRE Trans. on Circuit Theory, Vol. CT-6, No. 1, pp. 5-12, March 1959.

A NOTE ON THE AUTONOMOUS BEHAVIOR OF FINITE STATE MACHINES

A. MICZO
IBM Corporation, Owego, New York

L. D. RUDOLPH
Systems and Information Science
Syracuse University, Syracuse, New York

Although a great deal is known about the relationships between the structure of an autonomous finite state machine and the algebraic properties of its output sequences, it appears that very little is known about the relationships between machine structure and the symbol occurrence rate, or weight distribution, of its outputs. The purpose of this note is to point out a connection between the number of output symbols that must be observed in order to determine the initial state of a machine and the number of low-weight sequences a machine is capable of producing. For simplicity, we restrict our attention to machines with the two symbol output alphabet $\Gamma = \{a,b\}$. The major result is that an L-step observable autonomous machine is capable of producing at most $2^{LH(\delta_a)+1}$ sequences in which the occurrence rate of "a" is δ_a or less for $0 < \delta_a \leq \frac{1}{3}$, where $H(x) = -x\log_2 x-(1-x)\log_2(1-x)$ is the binary entropy function.

REALIZATION OF THRESHOLD SEQUENTIAL MACHINES BASED ON A MODIFICATION OF THE THEORY OF COMPLETE MONOTONICITY

Samuel G. Burgiss
Terminal Communications, Inc.
Raleigh, North Carolina

John Staudhammer
North Carolina State University
Raleigh, North Carolina

ABSTRACT

The theory of complete monotonicity is used to obtain the theory and procedure for the realization of threshold sequential machines. Reductions are made in this expansion to increase the efficiency of the realization. Algorithms are presented which allow the derivation of possible state, input, and output partitions for single threshold gate realization. Implementation of the procedure is possible using only very small amounts of computer storage and requiring very little time for execution.

INTRODUCTION

The research described provides a system of automated threshold logic design of a certain class of logic which greatly reduces engineering time and cost. This work provides an efficient automated method of designing nonautonomous sequential machines. Algorithms are presented which determine a state assignment for realizing a machine in one level of threshold gates. These algorithms can be applied in an automated manner or manually. When used in an automated manner, the algorithms can be segmented to be used on very small digital computers. This flexibility allows the designer to apply the algorithms using available facilities.

Early work in the application of threshold logic to the synthesis of sequential machines was done by Masters [1]. His dissertation presents a study based on a model having a single level of threshold gates in the combinational section of the sequential machine and delay elements in the memory section. The greatest contribution of this work is on autonomous machines. Hadlock [2] used the same sequential machine model. This study is based on the use of the condition of 2-asummability as a necessary and sufficient condition to prove that a function of seven or fewer variables is realizable with a single threshold gate. The condition of 2-asummability is rigorously defined and used in an algorithm to develop tables from the next-state and output functions of a nonautonomous synchronous sequential machine. Several lists are produced from the tables and finally a function is generated from which a set of assignment functions for input, output and state variables may be chosen. Methods are applied to reduce the problem, but application of the algorithm is quite laborious.

The problem of automated synthesis of nonautonomous sequential machines warrants further investigation. It is the intent of this paper to present a method of automated synthesis for nonautonomous sequential machines which can be efficiently applied on almost any computer.

Complete monotonicity was studied by R. O. Winder [3] in 1962. He showed that complete monotonicity is a necessary and sufficient condition to prove linear separability (i.e., realizability with one threshold gate) for functions of six or fewer variables. This test was proven to apply to fewer than nine variables by the continued work of Winder [4] and Muroga, et al. [5].

APPLICATION OF THE THEORY OF COMPLETE MONOTONICITY

By defining a Boolean function

$$w = f(x_1, x_2, \ldots, x_n)$$

on an n-dimensional hypercube having 2^n vertices, a set of vectors $\{v_i\}_1^{2^n}$ can be established between the origin and each vertex p_i. Each p_i has a value which is represented by

$$f(p_i) = 0$$

or

$$f(p_i) = 1.$$

Definition I. [3] If a Boolean function has a set of j not-necessarily-distinct vectors $\{v_i\}$ such that

$$f(p_i^1) = 1$$

and another set of j not-necessarily-distinct vectors $\{v_i^2\}$ such that

$$f(p_i^2) = 0$$

and

$$\sum_{i=1}^{j} v_i^1 = \sum_{i=1}^{j} v_i^2$$

then the function w is j-summable. Elgot [6] has shown that a function which is completely monotonic is not 2-summable and vice versa. This leads to the following theorem which can be proven from the above definition and Elgot's work.

Theorem I. [7] If and only if for a Boolean function w_1 all pairs of sets of two not-necessarily-distinct vertices (p_i), which are represented by

$$|\{p_i^1\}_1^2 : \{p_i^2\}_1^2|$$

where

$$f(p_i^1) = 1$$

and

$$f(p_i^2) = 0$$

have vectors related by

$$v_1^1 + v_2^1 \neq v_1^2 + v_2^2 \quad ;$$

then the function w is completely monotonic.

Theorem I provides a test for complete monotonicity which can be applied to both the next-state variable function (F^σ) and the output variable function (F^θ). For the next-state function each of the pairs of sets

$$\delta_2 = |\{{}_2\sigma_i^1\}_1^2 : \{{}_2\sigma_i^2\}_1^2|$$

must be investigated. Each of these pairs of sets is generated from a pair of sets

$$\beta = |\{\iota_i^1\}_1^2 : \{\iota_i^2\}_1^2|$$

of the input set and a pair of sets

$$\alpha = |\{{}_1\sigma_i^1\}_1^2 : \{{}_1\sigma_i^2\}_1^2|$$

of the state set which are combined to produce the pair of sets

$$\delta_1 = |\{\iota_i^1, {}_1\sigma_i^1\}_1^2 : \{\iota_i^2, {}_1\sigma_i^2\}_1^2|$$

Each δ_1 yields a δ_2 by the application of the next-state function F^σ. When a state assignment is determined, δ_1, δ_2, σ, ι, and F^σ are represented by $\tilde{\delta}_1$, $\tilde{\delta}_2$, $\tilde{\sigma}$, $\tilde{\iota}$, and $\tilde{F}^\sigma$ in their binary coded form.

When an assignment is applied to some types of β's and γ's to produce $\tilde{\beta}$'s and $\tilde{\gamma}$'s, the corresponding $\tilde{\delta}_1$'s cannot have elements related by

$$\tilde{\iota}^1_1, {}_1\tilde{\sigma}^1_1 + \tilde{\iota}^1_2, {}_1\tilde{\sigma}^1_2 = \tilde{\iota}^2_1, {}_1\tilde{\sigma}^2_1 + \tilde{\iota}^2_2, {}_1\tilde{\sigma}^2_2 \ .$$

If these types are eliminated, the testing of pairs of sets to satisfy Theorem I is decreased. It is evident that a $\tilde{\delta}_1$ has elements related by

$$\tilde{\iota}^1_1, {}_1\tilde{\sigma}^1_1 + \tilde{\iota}^1_2, {}_1\tilde{\sigma}^1_2 = \tilde{\iota}^2_1, {}_1\tilde{\sigma}^2_1 + \tilde{\iota}^2_2, {}_1\tilde{\sigma}^2_2$$

if and only if the corresponding $\tilde{\beta}$ has elements related by

$$\tilde{\iota}^1_1 + \tilde{\iota}^1_2 = \tilde{\iota}^2_1 + \tilde{\iota}^2_2$$

and the corresponding $\tilde{\gamma}$ has elements related by

$${}_1\tilde{\sigma}^1_1 + {}_1\tilde{\sigma}^1_2 = {}_1\tilde{\sigma}^2_1 + {}_1\tilde{\sigma}^2_2 \ .$$

It is necessary to determine what types of β's and γ's produce $\tilde{\beta}$'s and $\tilde{\gamma}$'s respectively, which satisfy this requirement.

Theorem II. [7] A necessary and sufficient condition that there exists an assignment α^ι for ι such that the relationship

$$\tilde{\iota}^1_1 + \tilde{\iota}^2_2 = \tilde{\iota}^2_1 + \tilde{\iota}^2_2$$

exists between the elements of $\tilde{\beta}$ is that either $\iota^1_1 = \iota^2_1 \neq \iota^1_2 = \iota^2_2$ or $\iota^1_1 = \iota^2_2 \neq \iota^1_2 = \iota^2_1$; $\iota^1_1 \neq \iota^1_2 \neq \iota^2_1 \neq \iota^2_2$ and $\iota^1_2 \neq \iota^2_2 \neq \iota^1_1 \neq \iota^2_1$; or $\iota^1_1 = \iota^1_2 = \iota^2_1 = \iota^2_2$ are true for β.

Another theorem can be stated for the γ's.

Theorem III. [7] A necessary and sufficient condition that there exists an assignment α^σ for σ such that the relationship

$${}_1\tilde{\sigma}^1_1 + {}_1\tilde{\sigma}^1_2 = {}_1\tilde{\sigma}^2_1 + {}_1\tilde{\sigma}^2_2$$

exists between the elements of $\tilde{\gamma}$ is that in γ either ${}_1\sigma^1_1 = {}_1\sigma^2_1 \neq {}_1\sigma^1_2 = {}_1\sigma^2_2$ or ${}_1\sigma^1_1 = {}_1\sigma^2_2 \neq {}_1\sigma^1_2 = {}_1\sigma^2_1$; ${}_1\sigma^1_1 \neq {}_1\sigma^1_2 \neq {}_1\sigma^2_1 \neq {}_1\sigma^1_2$ and ${}_1\sigma^1_2 \neq {}_1\sigma^2_2 \neq {}_1\sigma^1_1 \neq {}_1\sigma^2_1$; or ${}_1\sigma^1_1 = {}_1\sigma^1_2 = {}_1\sigma^2_1 = {}_1\sigma^2_2$ are true.

The conditions of Theorem II and Theorem III are complex, but they can be reduced to a simpler notation for application to pairs of sets of both ι and σ. The desired relationships between the elements of a pair of sets

$$|a,b : c,d|$$

can be defined as follows:

Type I.

$a \neq b \neq c \neq d$ and $b \neq d \neq a \neq c$

Type II.

$a = c \neq b = d$ or $a = d \neq b = c$

Type III.

$a = b = c = d$

In the following the term 'irredundant' pair is used. A pair of sets $|\{p^1_i\}^2_1 : \{p^2_i\}^2_1|$ is irredundant if $p^1_i \neq p^2_i$ for all possible pairs of p^1_i and p^2_i. A pair of sets $\tilde{\delta}_1$ will satisfy the conditions contained in the above theorems if the corresponding δ_2 is irredundant.

Of the nine possible combinations of β with respect to types of γ, the following can yield irredundant δ_2's [7]:

$$\beta_I\gamma_I, \beta_I\gamma_{II}, \beta_I\gamma_{III}, \beta_{II}\gamma_I, \beta_{II}\gamma_{II}, \beta_{III}\gamma_I$$

The preceeding theory also applies to the output function of a sequential machine. For this function, the pairs of sets

$$\eta_2 = |\{\theta_i^1\}_1^2 : \{\theta_i^2\}_1^2|$$

must be investigated. Each of these pairs of sets is generated from a β pair of sets and a γ pair of sets. The six combinations of types of β and γ stated previously are the combinations which can yield irredundant η_2's. The function F^θ for the machine to be designed must be used to determine if such η_2's are actually irredundant.

Irredundant δ_2 and η_2 pairs of sets are generated by enumerating all β and γ pair of sets combinations from F^σ and F^θ which belong to the six possibly irredundant types of β and γ pairs of sets and by testing each of these pairs of sets to obtain those which are irredundant. The number of elements which must be tested in the application of complete monotonicity as previously stated is

$$\Sigma_{n_1 n_2} = 8(S_{n_1} \cdot n_2 + n_1 \cdot S_{n_2} + B_{n_1} \cdot B_{n_2}) + 32(S_{n_1} \cdot B_{n_2} + B_{n_1} \cdot S_{n_2})$$
$$+ 64(S_{n_1} \cdot S_{n_2})$$

where n_1 is the order of ι; n_2 is that of σ; $B_k = \binom{k}{2}$; and $S_k = \sum_{i=1}^{k-2} (i+1)\binom{i}{2}$.

For a machine of eight inputs and sixteen states $\Sigma_{n_1 n_2}$ equals $794\text{x}10^5$; for sixteen inputs and sixteen states, this quantity is $195\text{x}10^7$. The expression for $\Sigma_{n_1 n_2}$ and the example values illustrate that the direct application of complete monotonicity is prohibitively laborious. Two algorithms modify the application of complete monotonicity such that it is relatively efficient.

THE ALGORITHMS

The first algorithm obtains the irredundant representative δ_2 and η_2 pairs of sets of F^σ and F^θ which have corresponding δ_1 and η_1 pairs of sets of the type $\beta_{II}\gamma_{II}$.

Algorithm I

Step 1: A=number of rows in F^σ; 2: B=number of columns in F^σ; 3: I=1; 4: J=1; 5: Test all elements in row I with column subscripts M such that $(J+1)\geq M\geq B$ and compare with element F^σ_{IJ}. Make a list ν whose length N is represented by $N\leq B$ and whose elements are column labels of element $F^\sigma_{IM}\neq F^\sigma_{IJ}$; 6: If there are no elements in ν, go to Step 16; 7: K=I+1; 8: If $F^\sigma_{KJ}= F^\sigma_{IJ}$, go to Step 14; 9: C=1; 10: L=C-th element of ν; 11: If $F^\sigma_{KJ}=F^\sigma_{KL}$ or $F^\sigma_{IL}=F^\sigma_{KL}$, go to Step 13; 12: If there are no pairs of sets in List A^σ or if there is no pair of sets in A^σ which is a representative pair of sets for $\delta_2=|F^\sigma_{IJ},F^\sigma_{KL};F^\sigma_{KJ},F^\sigma_{IL}|$, append the representative form of δ_2 to List A^σ; 13: If C≠N, C=C+1 and go to Step 10; 14: If K≠A, K=K+1 and go to Step 8; 15: J=J+1 and if J≠B go to Step 5; 16: I=I+1 and if I≠A go to Step 4; 17: Repeat

Steps 1 through 16 replacing σ with θ and δ_2 with η_2.

The second algorithm selects irredundant representative pairs of sets from F^{σ} and F^{θ} from which sets of partitions for input, state and output assignments may be made.

Algorithm II

Step 1: A=number of rows in F^{σ}; 2: B=number of columns in F^{σ}; 3: I=1; 4: J=1; 5: C=0; 6: N=J+1; 7: If $F^{\sigma}_{IN}=F^{\sigma}_{IJ}$, go to Step 22; 8: C=1; 9: K=I+1; 10: L=1; 11: If L=J or L=N, go to Step 20; 12: If $F^{\sigma}_{KL}=F^{\sigma}_{IN}$, go to Step 20; 13: P=1; 14: If P=J or P=N or P=L, go to Step 19; 15: If $F^{\sigma}_{KL}=F^{\sigma}_{KP}$ or $F^{\sigma}_{KP}=F^{\sigma}_{IJ}$, go to Step 19; 16: Convert the pair of sets $|F^{\sigma}_{IJ},F^{\sigma}_{KL} : F^{\sigma}_{IN},F^{\sigma}_{KP}|$ into its representative pair of sets and call it $\tilde{\delta}_2$; 17: Convert the pair of sets $|J,L : N,P|$ to its representative pair of sets and call it $\tilde{\beta}$; 18: If there are no entries in List B^{σ}_1 or if there is no entry which has a δ_2 equivalent to $\tilde{\delta}_2$ and a β equivalent to $\tilde{\beta}$, enter $\tilde{\delta}_2$ and $\tilde{\beta}$ in List B^{σ}_1 as a δ_2 and β entry; 19: If P≠B, P is incremented by 1 and go to Step 14; 20: If L≠B, L=L+1 and go to Step 11; 21: If K≠A, K=K+1 and go to Step 10; 22: If N≠B, N=N+1 and go to Step 7; 23: If C=0, go to Step 26; 24: J=J+1; 25: If J≠B, go to Step 5; 26: I=I+1; 27: If I≠A, go to Step 4; 28: Repeat Steps 1 through 27 replacing σ with θ and δ_2 with η_2; 29: A=number of rows in F^{σ}; 30: B=number of columns in F^{σ}; 31: I=1; 32: J=1; 33: L=J+1; 34: If $F^{\sigma}_{IL}=F^{\sigma}_{IJ}$, go to Step 46; 35: K=J+1; 36: If K=L, go to Step 45; 37: If $F^{\sigma}_{IK}=F^{\sigma}_{IL}$, go to Step 45; 38: M=L+1; 39: If M=K, go to Step 44; 40: If $F^{\sigma}_{IM}=F^{\sigma}_{IJ}$ or $F^{\sigma}_{IM}=F^{\sigma}_{IK}$, go to Step 44; 41: Convert the pair of sets $|F^{\sigma}_{IJ},F^{\sigma}_{IK}:F^{\sigma}_{IL},F^{\sigma}_{IM}|$ into its representative pair of sets and call it δ_2; 42: Convert the pair of sets $|J,K:L,M|$ into its representative pair of sets and call it $\tilde{\beta}$; 43: If there are no entries in List B^{σ}_1 or if there is no entry which has a δ_2 equivalent to $\tilde{\delta}_2$ and a β equivalent to $\tilde{\beta}$, enter $\tilde{\delta}_2$ and β in the List B^{σ}_1 as a $\tilde{\delta}_2$ and β entry; 44: If M≠B, M=M+1 and go to Step 39; 45: If K≠B, K=K+1 and go to Step 36; 46: L=L+1; 47: If L≠B, go to Step 34; 48: If J≠(B-3), J=J+1 and go to Step 33; 49: If I≠A, I=I+1 and go to Step 32; 50: Repeat Steps 29 through 49 replacing σ with θ and δ_2 with η_2; 51: A=number of rows in F^{σ}; 52: B=number of columns in F^{σ}; 53: I=1; 54: J=1, 55: C=0; 56: N=J+1; 57: If $F_{NI}=F_{JI}$, go to Step 72; 58: C=1; 59: K=I+1; 60: L=1; 61: If L=J or L=N, go to Step 70; 62: If $F^{\sigma}_{LK}=F^{\sigma}_{NI}$, go to Step 70; 63: P=1; 64: If P=J or P=N or P=L, go to Step 69; 65: If $F^{\sigma}_{PK}=F^{\sigma}_{LK}$ or $F^{\sigma}_{PK}=F^{\sigma}_{JI}$, go to Step 69; 66: Convert the pair of sets $|F^{\sigma}_{JI},F^{\sigma}_{LK}:F^{\sigma}_{NI},F^{\sigma}_{PK}|$ into its representative pair of sets and call it $\tilde{\delta}_2$; 67: Convert the pair of sets $|J,L:N,P|$ into its representative pair of sets and call it $\tilde{\gamma}$; 68: If there are no entries in List B^{σ}_2 or if there is no entry which has a δ_2 equivalent to $\tilde{\delta}_2$ and a γ equivalent to $\tilde{\gamma}$, enter $\tilde{\delta}_2$ and $\tilde{\gamma}$ in List B^{σ}_2 as a δ_2 and γ entry; 69: If P≠B, P=P+1 and go to Step 64; 70: If L≠B, L=L+1 and go to Step 61; 71: If K≠A, K=K+1 and go to Step 60; 72: If N≠B, N=N+1 and go to Step 57; 73: If C=0, go to Step 76; 74: J=J+1; 75: If J≠B, go to Step 55; 76: I=I+1; 77: If I≠A, go to Step 54; 78: Repeat Steps 51 through 77 replacing σ with θ and δ_2 with η_2; 79: A=number of rows in F^{σ}; 80: B=number of columns in F^{σ}; 81: I=1; 82: J=1; 83: L=J+1; 84: If $F^{\sigma}_{LI}=F^{\sigma}_{JI}$, go to Step 96; 85: K=J+1; 86: If K=L, go to Step 95; 87: If $F^{\sigma}_{KI}=F^{\sigma}_{LI}$, go to 95; 88: M=L+1; 89: If M=K, go to Step 94; 90: If $F^{\sigma}_{MI}=F^{\sigma}_{JI}$ or $F^{\sigma}_{MI}=F^{\sigma}_{IK}$, go to Step 94; 91: Convert the pair of sets $|F^{\sigma}_{JI},F^{\sigma}_{KI}:F^{\sigma}_{LI},F^{\sigma}_{MI}|$ to its representative pair of sets and call it $\tilde{\delta}_2$; 92: Convert the pair of sets |J,K:L,M| to its representative pair of sets and call it $\tilde{\gamma}$; 93: If there are no entries in List B^{σ}_2 or if there is no entry which has a δ_2 equivalent to $\tilde{\delta}_2$ and a γ equivalent to $\tilde{\gamma}$, enter $\tilde{\delta}_2$ and $\tilde{\gamma}$ in List B^{σ}_2 as a δ_2 and γ entry; 94: If M≠B, M=M+1 and go to Step 89; 95: If K≠B, K=K+1 and go to Step 86; 96: L=L+1; 97: If L≠B, go to Step 84; 98: If J≠(B-3), J=J+1 and go to Step 83; 99: If I≠A, I=I+1 and go to Step 82; 100: Repeat Steps 79 through 99 replacing σ with θ and δ_2 with η_2; 101: A=number of rows in F^{σ}; 102: B=number of columns in F^{σ}; 103: I=1; 104: J=1; 105: M=I+1; 106: N=1; 107: If N=J, go to Step 109; 108: If $F^{\sigma}_{IJ}=F^{\sigma}_{MN}$, go to Step 129; 109: K=I+1; 110: If K=M, go to Step 128; 111: L=1; 112: If L=J or L=N, go to Step 127; 113: If $F^{\sigma}_{KL}=F^{\sigma}_{MN}$, go to Step 127; 114: G=I+1; 115: If G=M or G=K, go to Step 126; 116: H=1; 117: If H=J or H=N or H=L, go to Step 125; 118: If $F^{\sigma}_{GH}= F^{\sigma}_{IJ}$ or $F^{\sigma}_{GH}=F^{\sigma}_{KL}$, go to Step 125; 119: Convert the pair of sets $|F^{\sigma}_{IJ},F^{\sigma}_{KL}:F^{\sigma}_{MN},F^{\sigma}_{GH}|$ into its representative pair of sets and call it $\tilde{\delta}_2$; 120: Convert the pair of sets |J,L:N,H| into its representative pair of sets and call it $\tilde{\beta}$; 121: Convert the pair of sets |I,K:M,G| into its representative pair of sets and call it $\tilde{\gamma}$; 122: If there are no entries in List B^{σ}_3, go to Step 124; 123: If there is an entry in List B^{σ}_3 which has a δ_2 equivalent to $\tilde{\delta}_2$ and a β equivalent to $\tilde{\beta}$, and a γ equivalent to γ, go to Step 125; 124: If there is no entry in List B^{σ}_1 which has a δ_2 equivalent to $\tilde{\delta}_2$ and a β equivalent to $\tilde{\beta}$, and there is no entry in List B^{σ}_2 which has a δ_2 equivalent to $\tilde{\delta}_2$ and a β equivalent to $\tilde{\beta}$, enter $\tilde{\delta}_2$, $\tilde{\beta}$ and $\tilde{\gamma}$ in List B^{σ}_3 as a δ_2, β and γ entry; 125: If H≠B, H=H+1 and go to Step 117; 126: If G≠A, G=G+1 and go to Step 115; 127: If L≠B, L=L+1 and go to Step 112; 128: If K≠A, K=K+1 and go to Step 112; 129: If N≠B, N=N+1 and go to Step 107; 130: If M≠A, M=M+1 and go to Step 106; 131: If J≠B, J=J+1 and go to Step 105; 132: If I≠(A-4), I=I+1 and go to Step 104; 133: Combine Lists B^{σ}_1, B^{σ}_2 and B^{σ}_3 to make up the list B^{σ} (when entries are inserted into List B^{σ}, the γ pair of sets should be empty for an entry from List B^{σ}_1 and the β pair of sets should be empty for an entry from List B^{σ}_2); 134: Repeat Steps 101 through 133 replacing σ with θ and δ_2 with η_2.

Note that a representative pair of sets |a,b:c,d| are related by $a \le b$, $c \le d$, and $a \le c$.

The first step in the synthesis of a sequential machine is the application of Algorithm I and Algorithm II to the next-state and output functions for the machine. This results in List A, composed of List A^{σ} and List A^{θ}, and List B, composed of List B^{σ} and List B^{θ}. List A has entries of δ_2 and and η_2 pairs of sets, and List B has a δ_2 or an η_2 pair of sets and a β

and/or a γ pair of sets for each entry. The above theorems are applied to List A and List B to yield the partition set

$$\pi = \{\psi, \Phi, x\}$$

composed of a set of input partitions (ψ), a set of state partitions (Φ), and a set of output partitions (x). Using these partitions, assignments are selected to yield coded input, state and output sets. These assignments result in coded next-state and output functions. From these functions, the next-state and output variable functions are written. Each of these functions is linearly separable and is realized with a single threshold gate.

IMPLEMENTATION

The realization procedure begins with the next-state and output functions of a sequential machine. Algorithm I and Algorithm II are applied to these functions to obtain List A and List B. The application of these algorithms can be implemented as a unit or can be efficiently segmented into as many as twelve parts to enable implementation on small computers (i.e., 4000 word minicomputers). The data obtained in either case consists of reduced lists of irredundant pairs of sets. Obtaining these lists without the algorithms precludes a practical application of complete monotonicity. The algorithms have been tested for validity and have been shown to be efficient. A test program (with diagnostics) has been implemented on an IBM 360 Model 75 (time shared system) computer. This program required seven-tenths of a second of processing time for examples having ι of order four and σ of order four. Core required (including diagnostics and I/O) was 37^k bytes; table size required for a 32 state and 32 input machine is 4512 bytes.

SUMMARY

An efficient method of automated realization for nonautonomous sequential machines has been obtained through the application of the theory of complete monotonicity. The basic theory has been expanded to include sequential machines and was shown to be prohibitively laborious. Extensive modifications were made in the form of Algorithm I and Algorithm II. The implementation of these algorithms has been shown to be efficient and applicable on most computer systems.

REFERENCES

1. Masters, G. M., Threshold logic synthesis of sequential machines, Ph.D. Dissertation, Stanford University, 1966.
2. Hadlock, F. O. and C. L. Coates, Realization of sequential machines with threshold elements, Laboratories for Electronics and Related Sciences, College of Engineering, The University of Texas, Austin, Texas, 1966.
3. Winder, R. O., Threshold logic, Ph.D. Dissertation, Department of Mathematics, Princeton University, 1962.
4. Winder, R. O., Enumeration of seven-argument threshold functions, IEEE Transactions on Electronic Computers, EC-14, 315-325 (1965).
5. Muroga, S., T. Tsuboi and C. R. Baugh, Enumeration of threshold functions of eight variables, IEEE Transactions on Computer, C-19, 818-825 (1970).
6. Elgot, C. C., Truth functions realizable by single threshold organs, IEEE Proceedings on Switching Theory and Logical Design, 225-245 (1960).

7. Burgiss, S. G., On the realization of threshold sequential machines, Ph.D. Dissertation, Department of Electrical Engineering, North Carolina State University, Raleigh, North Carolina (1971).

ELIMINATION OF SPURIOUS OUTPUT PULSES FOR MOORE TYPE SEQUENTIAL MACHINES

JON G. BREDESON AND PAUL T. HULINA
Department of Electrical Engineering
Pennsylvania State University
University Park, Pennsylvania

ABSTRACT

This paper deals with elimination of spurious pulses on the output functions of a synchronous sequential machine of the Moore type. Two state assignment methods are given to code the secondary states to avoid static and dynamic function hazards on the outputs. The output networks can also be realized to avoid static and dynamic logic hazards.

INTRODUCTION

There are many applications when synchronous sequential machines must be designed to avoid spurious pulse on the outputs. In certain applications a momentary "1" pulse, "0" pulse or "0 - 1 - 0 - 1" transition cannot be tolerated [1,2,3]. One example would be time slot decoding for multiplex applications, i.e., n states and n output functions where one and only one function equals "1" for each state. Momentary "1" pulses cannot be tolerated for multiplex applications [4].

In the normal "real world" of design, synchronous sequential circuits are implemented using flip-flops. The output networks which are a function of the state only (Moore type) use as inputs the outputs of the flip-flops. The output networks which are a function of both the input and state (Mealy type) normally are "ANDED" with the clock to produce a pulse output [1]. The clock pulses are connected to each flip-flop and excite each flip-flop simultaneously; however, the outputs of the flip-flops do not change simultaneously. Thus, the output networks of the machine can see multiple input changes which do not occur simultaneously. The same situation occurs if delay elements are used to synthesize the machine.

We intend to show that by choosing an appropriate state assignment and a proper output network design, spurious pulses on the outputs can be eliminated.

FUNCTION HAZARDS

In this paper, the input state for the network output functions will actually be the state of the machine. An input state for the output functions will be assumed to be a function of an n-tuple $(y_1, y_2, \ldots, y_n)$. To insure no spurious pulses are present, we must first insure no function hazards exist. Some definitions are in order.

Definition 1: Assume $A_1, A_2, \ldots, A_m$ have the same logic values for $y_1, y_2, \ldots, y_p$ $(p \leq n)$. The subcube of inputs states $A_1, A_2, \ldots, A_m$ is the set of all input states with the same logic values for $y_1, y_2, \ldots, y_p$ variables. $A\varepsilon(A_1, A_2, \ldots A_m)$ denotes A is in the subcube of $A_1, A_2, \ldots, A_m$.

As an example of the definition consider the following input states.

$A_1 = (1, 0, 0, 1)$ $A_2 = (0, 0, 0, 1)$

$A_3 = (0, 0, 0, 0)$ $A = (1, 0, 0, 0)$

By the definition

$$A\varepsilon(A_1, A_2, A_3), \quad A\varepsilon(A_1, A_3)$$

but

$$A\not\varepsilon(A_1, A_2).$$

The following definition is equivalent to Eichelberger's [5] definition of a function hazard.

Definition 2: A Boolean function f contains a static function hazard for the input change A to C if and only if

(1) $f(A) = f(C)$

(2) There exists at least one input state $B\varepsilon(A, C)$ such that $f(B) \neq f(A)$.

Definition 3: A Boolean function f contains a dynamic function hazard for the input change A to D if and only if

(1) $f(A) \neq f(D)$

(2) There exists at least one pair of inputs states B and C such that

a. $B\varepsilon(A, D)$ and $C\varepsilon(B, D)$ and

b. $f(B) = f(D)$ and $f(C) = f(A)$

Examples of static and dynamic function hazards are given in Figure 1 for input changes A to C and A to D, respectively.

y_3 \ $y_1 y_2$	00	01	11	10
0	0^A	1^B	1	0
1	0	0^C	1^D	1

$f(y_1, y_2, y_3)$

$A = (0, 0, 0)$
$B = (0, 1, 0)$
$C = (0, 1, 1)$
$D = (1, 1, 1)$

Figure 1 - Example of Function Hazards

It should be apparent that if a function hazard exists (either static or dynamic), a sequence of input changes can exist where a function may have a multiple (more than one) output change. Then it is possible that delays in the network realizing a function with function hazards could cause a multiple change in the output of the network. Thus the following theorem:

Theorem 1: If a Boolean function f contains a static or dynamic function hazard for an input change A to B, it is impossible to realize a network for f which will eliminate possible multiple output changes for input change A to B.

In the next section, we will give a method of assigning secondary states to insure no function hazards exist on the output functions. The method is based on partitioning of states.

THE STATE ASSIGNMENT PROBLEM

In this paper, we will assume synchronous machine operation and output functions that are dependent only on the states of the machine (Moore type) [6]. The states in the flow table will be designated as q_i, the secondary state assigned to q_i as s_i (written $q_i \leftrightarrow s_i$), and the output function as Z_o. The transition from state q_i to state q_j will be presented as $q_i \rightarrow q_j$. Since synchronous machine operation is assumed races will not affect proper sequencing of states.

We must assign secondary states to each state in the flow table to avoid both static and dynamic function hazards on the output functions. Assume n internal variables are needed so each internal state is represented by the n-tuple $s = (y_1, y_2, \ldots, y_n)$. From Z_o a new output function Z_o^* must be formed. Let $Z_o^*(s_i) = Z_o(q_i)$ for all $s_i \leftrightarrow q_i$. The new function Z_o^* must be defined on the remaining secondary states which do not correspond to specified states in the flow table in a manner to avoid static and dynamic function hazards.

Consider state transition $q_i \rightarrow q_j$ for a Moore machine. If $Z_o(q_i) = Z_o(q_j)$, then to avoid a secondary state assignment which causes a static function hazard, we must insure there is no $s \varepsilon (s_i, s_j)$ such that the function $Z_o^*(s) \neq Z_o^*(s_j) = Z_o(q_j)$. Note that s may or may not correspond to a state q_k in the machine. If $Z_o(q_i) \neq Z_o(q_j)$ then to avoid a dynamic function hazard we must insure for all $s \varepsilon (s_i, s_j)$, where $Z_o^*(s) = Z_o^*(s_j) = Z_o(q_j)$, that no static function hazard occurs on function Z_o^* in the transition $s \rightarrow s_j$ [7]. It is also possible to show that if the function Z_o^* contains no static function hazard for any transition where the output starts and remains at logical 1, (or 0) then Z_o^* cannot contain any dynamic function hazards [7].

A procedure will now be given for defining Z_o^* such that there will be no static function hazards for input changes s_i

to s_j ($q_i \rightarrow q_j$) when $Z_o^*(s_j) = 0$, and no static function hazards for all input changes s to s^* when $Z_o^*(s) = Z_o^*(s^*) = 1$.

The following definition by Tracey [8] will be needed for what is to follow:

Definition 4: A partition π on a set S is a collection of subsets of S such that their pairwise intersection is the null set. The disjoint subsets are called the blocks of π. If the set union of these subsets is S, the partition is completely specified; otherwise, the partition is incompletely specified. Elements of S that do not appear in it are called unspecified or optional elements with respect to that partition. A y-variable y_i induces the partition $\pi = \{s_1, s_2, s_m; s_{m+1}, \ldots, s_k\}$ if the value of y_i is "0" for $s_1, s_2, \ldots, s_m$ and "1" for $s_{m+1}, \ldots, s_k$ or vice versa.

Theorem 2: There exists no input state s in both subcube $(s_1, s_2, \ldots, s_m)$ and subcube $(s_{m+1}, s_{m+2}, \ldots, s_k)$ if and only if there exists at least one y-variable which partitions the set of input states $\{s_1, s_2, \ldots, s_m\}$ and the set of input states $\{s_{m+1}, \ldots, s_k\}$ into separate blocks.

Proof: Assume the variable y_i partitions the two sets of input states into separate blocks. Then the y_i variable of $s_1, s_2, \ldots, s_m$ is equal to say x and the y_i variable of $s_{m+1}, s_{m+2}, \ldots, s_k$ is equal to $\bar{x}$. If $s \varepsilon (s_1, s_2, \ldots, s_m)$ and $s \varepsilon (s_{m+1}, \ldots, s_k)$ implies that the y_i variable of s must equal both x and $\bar{x}$. This is clearly impossible.

Assume there exists no input variable which partitions the two sets of input states. An input state s which is in the subcube of both sets of states will be constructed. For any y_i variable which has the same value for all states in either set let the y_i variable of s have the same value. (Note that both sets will not have conflicting values of any y_i). The remaining variables may be either "0" or "1". s constructed in this way is in both subcube $(s_1, s_2, \ldots, s_m)$ and subcube $(s_{m+1}, \ldots, s_k)$. Hence, the theorem.

A state assignment method which uses the results of Theorem 4 will be presented, and an output function which depends on the secondary states will be determined so as to eliminate all possible function hazards for allowable transitions.

For each output function Z_o determine all states q_i such that $Z_o(q_i) = 1$. Assume for convenience they are $q_1, q_2, \ldots, q_m$. Then find all transitions $q_i \rightarrow q_j$ where $Z_o(q_i) = Z_o(q_j) = 0$, and all remaining states q_k where $Z_o(q_k) = 0$. Ignore states with unspecified outputs.

Definition 5: The *partition list* includes for each Z the collection of partitions of the type $\{\overline{q_1, q_2, \ldots, q_m}; \overline{q_i, q_j}\}$ and $\{\overline{q_1, q_2, \ldots, q_m}; \overline{q_k}\}$ where the states are defined as in the previous paragraph. Also include all partitions of the type $\pi = \{\overline{q}_i; \overline{q}_j\}$ where $q_i \neq q_j$.

The problem of finding a suitable state assignment then becomes one of finding a minimum set of partitions (called τ- partitions) which cover all the partitions in the partition list. The procedure for finding these partitions is exactly the same as outlined by Tracey [8], and will not be given here.

The function Z_o^* which depends on the internal states s of the machine (input states for Z_o^*) is defined as follows ($q_i \leftrightarrow s_i$):

(1) If $s \varepsilon (s_1, s_2, \ldots, s_m)$, let $Z_o^*(s) = 1$

(2) If $s \not\varepsilon (s_1, s_2, \ldots, s_m)$, let $Z_o^*(s) = 0$

Theorem 3: The function Z_o^* will not contain any static or dynamic function hazards for any input change corresponding to transitions in the flow table. (i.e., if $q_i \rightarrow q_j$, then the input change s_i to s_j will not contain a function hazard on Z_o^*.)

Proof: First, it will be shown that no static function hazards occur for any input change s_i to s_j where $Z_o^*(s_i) = Z_o^*(s_j) = 0$, $s_i \leftrightarrow q_i$, $s_j \leftrightarrow q_j$, and $q_i \rightarrow q_j$. Since $\{\overline{q_1, q_2, \ldots, q_m}; \overline{q_i, q_j}\}$ occurs in the partition list, this implies (Theorem 2) there exists no s such that $s \varepsilon (s_1, s_2, \ldots, s_m)$ and $s \varepsilon (s_i, s_j)$. Thus, $Z_o^*(s) = 0$ for all $s \varepsilon (s_i, s_j)$; hence, a static function hazard does not exist. If q_k is not involved in a transition and $Z_o(q_k) = 0$, then $\{\overline{q_1, q_2, \ldots, q_m}; \overline{q_k}\}$ is in the partition list and by Theorem 2; $Z_o^*(s_k) = 0$.

Now if we can show no static function hazard exists for any transition $s \rightarrow s'$ where $Z_o^*(s) = Z_o^*(s') = 1$, then Z_o^* will contain no static or dynamic function hazards [7]. Let $s'' \varepsilon (s, s')$ and assume $Z_o^*(s'') = 0$. However, $s'' \varepsilon (s, s')$ implies $s'' \varepsilon (s_1, s_2, \ldots, s_m)$ and $Z_o^*(s'')$ is defined as 1. Thus, the theorem.

To illustrate the procedure an example will now be given. Consider machine B as given in Figure 2. The machine is a time slot decoder of 5 states.

Input (clock only)

	C = 0	C = 1	$Z_1 Z_2 Z_3 Z_4 Z_5$
States 1	1	2	1 0 0 0 0
2	2	3	0 1 0 0 0
3	3	4	0 0 1 0 0
4	4	5	0 0 0 1 0
5	5	1	0 0 0 0 1

Figure 2 - Machine B

The partition list will include the following partitions:

Z_1: $\pi_1 = \{\overline{1};\ \overline{2, 3}\}$ $\quad \pi_2 = \{\overline{1};\ \overline{3, 4}\}$ $\quad \pi_3 = \{\overline{1};\ \overline{4, 5}\}$

Z_2: $\pi_4 = \{\overline{2};\ \overline{3, 4}\}$ $\quad \pi_5 = \{\overline{2};\ \overline{4, 5}\}$ $\quad \pi_6 = \{\overline{2};\ \overline{5, 1}\}$

Z_3: $\pi_7 = \{\overline{3};\ \overline{4, 5}\}$ $\quad \pi_8 = \{\overline{3};\ \overline{5, 1}\}$ $\quad \pi_9 = \{\overline{3};\ \overline{1, 2}\}$

Z_4: $\pi_{10} = \{\overline{4};\ \overline{5, 1}\}$ $\quad \pi_{11} = \{\overline{4};\ \overline{1, 2}\}$ $\quad \pi_{12} = \{\overline{4};\ \overline{2, 3}\}$

Z_5: $\pi_{13} = \{\overline{5};\ \overline{1, 2}\}$ $\quad \pi_{14} = \{\overline{5};\ \overline{2, 3}\}$ $\quad \pi_{15} = \{\overline{5};\ \overline{3, 4}\}$

Since all partitions of the type $\pi = \{\overline{q_i};\ \overline{q_j}\}$, $q_i \neq q_j$ are covered by the set of partitions already listed, they need not be added to the partition list.

The following set of τ-partitions were found by Tracey's method and will cover all the partitions in the partition list. (Other sets of τ-partitions also exist):

$$\tau_1 = \{\overline{1, 2};\ \overline{3, 4, 5}\}$$

$$\tau_2 = \{\overline{1, 5};\ \overline{2, 3, 4}\}$$

$$\tau_3 = \{\overline{2, 3};\ \overline{1, 4, 5}\}$$

If y_1, y_2, and y_3 are used to code τ_1, τ_2, τ_3, respectively, the following state assignment results for machine B:

	y_1	y_2	y_3
1	0	0	1
2	0	1	0
3	1	1	0
4	1	1	1
5	1	0	1

The output function Z_1^* is given in Figure 3.

y_3 \ $y_1 y_2$	00	01	11	10
0	0	0^2	0^3	0
1	1^1	0	0^4	0^5

Figure 3 - Output function Z_1^* for Machine B

Obviously in the generation of the partitions, all of the states in which $Z_o(q_i) = 0$ could be grouped in one block and states q_i, where $Z_o(q_i) = 1$, or pairs of states in the other block. The same results will hold with "0" replacing "1" and vice versa. The best method (fewer y-variables) is to determine for each output function Z_i the number of states with "0" and with "1" outputs. Whichever is less group those states into one block and in the other block of the partitions include states of the other type.

ELIMINATION OF LOGIC HAZARDS

The following definitions will be used for elimination of hazards due to delays. Definition 6 is equivalent to the logic hazard definition of Eichelberger [5] or Howe and Coates [9].

Definition 6: A network contains a static logic hazard for the input change A to B if and only if

(1) $f(A) = f(B)$

(2) No static function hazard exists for change A to B.

(3) During the input change A to B, a spurious hazard pulse may be present on the output.

Definition 7: A network contains a dynamic logic hazard for the input change A to B if and only if

(1) $f(A) \neq f(B)$

(2) No dynamic function hazard exists for input change A to B.

(3) During the input change A to B, a momentary "0" output and a momentary "1" output may appear.

The state assignment method insures us no function hazards will be present. We now need to give a network realization which will be free of logic hazards. It has been shown that any network realized with one prime implicant will not contain any static or dynamic logic hazards [7]. Thus, we need only to show each Z_o^* contains only one prime implicant.

Theorem 7: The function Z_o^* as defined in the previous section will contain only one prime implicant.

Proof: Let $y_1, y_2, \ldots, y_k$ be the y-variables which have the same value for all $s \varepsilon (s_1, s_2, \ldots, s_m)$. Consider the literals $y_1^*, y_2^*, \ldots, y_k^*$ where $y_i^* = y_i$ if the y_i value of any $s \varepsilon (s_1, s_2, \ldots, s_m)$ is "1" and $y_1^* = \bar{y}_i$ otherwise. The assertion is $Z_i^* = y_1^* y_2^* \ldots, y_k^*$. For all input states $s \varepsilon (s_1, s_2, \ldots, s_m)$ obviously $y_1^* y_2^*, \ldots, y_k^* = 1 = Z_o^*(s)$. $s \not\varepsilon (s_1, s_2, \ldots, s_m)$ implies at least one y_i^* is not equal to "1" and hence $Z_o^*(s) = 0 = y_1^* y_2^*, \ldots, y_k^*$. Since Z_o^* equals one product term, then obviously Z_o^* contains one prime implicant.

SIMPLIFIED STATE ASSIGNMENT

The method presented previously for eliminating function hazards can be quite cumbersome as the size of the machine increases. The number of partitions in the partition list can increase at a substantial rate. However, there is a simplified method of eliminating function hazards.

Construct a partition list in the following manner. For each output function Z_o generate a partition π_o which includes all q_i such that $Z_o(q_i) = 0$ in one block and in the other block of π_o all q_j such that $Z_o(q_j) = 1$. If $Z_o(q_k)$ is undefined do not include it in π_o. Also, include all possible partitions of the type $\pi = \{\bar{q}_k; \bar{q}_j\}$ where $q_k \neq q_j$, and q_k and q_j are not in separate blocks of some partitions.

Find the set of τ-partitions to cover this partition list. Once the state assignment has been made, define Z_o^* on the secondary states in the same manner as the original procedure.

Theorem 8: The function Z_i^* for the simplified state assignment method will contain no static or dynamic function hazards.

Proof: Since the partitions in the partition list of the first method are obviously covered by the partitions in the new partition list and Z_o^* is defined as before, no function hazards will exist.

The realization of the output functions Z_o^* will contain no logic hazards since the results of this section hold. Furthermore, the output functions Z_o^* will now be a function of only one variable and will require no gates to implement.

The procedure is quite easy to apply and will usually give the fewest number of internal variables needed as long as the number of output functions is less than or equal to the smallest integer greater than or equal to $\log_2$ of the number of states.

CONCLUSION

Two methods were given for assigning states and defining the output function in a manner to avoid static and dynamic function hazards. The first method in most cases produces a state assignment utilizing the minimum number of internal variables. The second method requires very little computation and is idealy suited when the number of output functions is less than the required number of internal variables. In many cases, this requirement is met.

The output network realization was shown to be free of multiple output changes due to stray delays. Furthermore, the second method produces a state assignment in which each output function depends on only one internal variable; hence, no gates are needed to synthesize the output network.

REFERENCES

1. E. J. McCluskey, Introduction to the Theory of Switching Circuits, New York: McGraw-Hill, 1965.

2. G. A. Maley, and J. Earle, The Logic Design of Transistor Digital Computers, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.

3. D. A. Huffman, "The Design and Use of Hazard-Free Switching Networks," J. ACM, Vol. 4, pp. 47 - 62, January 1957.

4. R. S. Ledley, Digital Computer and Control Engineering, New York: McGraw-Hill, 1960.

5. E. B. Eichelberger, "Hazard Detection in Combinational and Sequential Switching Circuits," IEEE International Conf. Rec. on Switching Circuit Theory and Logical Design, Vol. 12, pp. 111 - 121, 1964.

6. E. F. Moore, "Gedanken-Experiments on Sequential Machines," in Automata Studies, Princeton, N. J.: Princeton University Press, 1956.

7. J. G. Bredeson and P. T. Hulina, "Elimination of Static and Dynamic Hazards in Combinational Switching Circuits," Proceedings 11th Annual Symp. on Switching and Automata Theory Conference, October, 1970.

8. J. H. Tracey, "The Internal State Assignment for Asynchronous Sequential Machines," IEEE Trans. Electronic Computers, Vol. EC-15, pp. 551 - 560, August 1966.

9. A. B. Howe, and C. L. Coates, "Logic Hazards in Threshold Networks," IEEE Trans. Computers, Vol. C-17, pp. 238 - 251, March 1968.

TRIGGER MACHINES: CANONICAL FORMS AND OTHER PROPERTIES

H. T. HAO
IBM Advanced Systems Development Division, Yorktown Heights, N.Y.

M. M. NEWBORN*
Department of Electrical Engineering and Computer Science,
Columbia University, New York, N.Y.

ABSTRACT

A synchronous sequential machine realizable in the form of a loop-free circuit consisting of trigger flip-flops and combinational gates is called a trigger machine. This paper studies the properties of the class of trigger machines. Most importantly, canonical circuit forms and canonical algebraic representations are presented. Other properties, many of which closely parallel those of definite machines, are also presented.

1. INTRODUCTION

This paper is a study of a class of (Moore type) synchronous sequential machines each realizable by a loop-free circuit consisting of trigger flip-flops and combinational gates. Machines in this class are called trigger machines and it is the purpose of this paper to outline their properties. Most importantly, canonical circuit forms and canonical algebraic representations are presented. Other properties, many of which closely parallel the properties of definite machines, are also presented. Some of the material included in this paper has been approached in a different manner earlier by Harlow and Coates[1] and Curtis[2]. Their work is credited where appropriate.

2. THE AOT MODULE AND ITS UNIVERSAL PROPERTY

With no loss of generality, we restrict our attention to machines with a single binary valued input x and a single binary valued output z. As usual a machine is described by a flow table. The initial state is circled. The set of states for which z is 1 is denoted as F. A machine is realized by a synchronous sequential circuit. The variable representing a signal at some point in a circuit is denoted as $[s_{i_1},s_{i_2},\cdots,s_{i_p}]$ where s_{i_j} (for $1 \leqslant j \leqslant p$) are those states of the circuit for which the variable is 1. The variable is said to contain states $s_{i_1},s_{i_2},\cdots,s_{i_p}$.

A circuit can be viewed as being composed of subcircuits or modules. A module is said to be universal[3] if for every machine there exists a corresponding circuit realization consisting of an interconnection of finite number of copies of that one module. Various modules have been shown to be universal--in particular, the AND-OR-DELAY (AOD) module[4,5]. In [6] it was shown that class of definite machines corresponds to the class of machines each realizable by a tree of AOD modules.

Definition 1. An AND-OR-TRIGGER (AOT) module is a two state module with inputs q, a and b, and output y having the input-output relation:

$$y_t = y_{t-1} \oplus (a_{t-1}\bar{q}_{t-1} + b_{t-1}q_{t-1}) \qquad \text{for } t \geqslant 1 \tag{1a}$$

$$y_0 = 0 \text{ or } 1 \tag{1b}$$

The module and its abbreviated circuit symbol are shown in Fig. 1. In the circuit realization of a machine shaded modules correspond to modules whose output are initially 1.

Theorem 1. The AOT module is a universal module.

---

*This work was supported in part by a National Science Foundation Grant NSF GK 14342.

Proof: This module can be shown to be output sufficient and thus universal using theorems presented by Ullman and Weiner (Theorems 3 and 4 of [3]). Curtis in a recent paper has also obtained a very similar result in that he has shown that every Moore model sequential machine can be realized by a polylinear sequential circuit using trigger flip-flop memory (Theorem 2 of [2]). Rather than presenting a formal proof, we outline an algorithm that yields an AOT circuit realization. In these circuits the input q to every module is connected to the input x of the machine; inputs a and b are connected to the outputs of other modules (including itself) or to logic constants. Note that we can write:

$$y_t = (y_{t-1} \oplus a_{t-1})\bar{x}_{t-1} + (y_{t-1} \oplus b_{t-1})x_{t-1} \qquad \textit{for } t \geqslant 1 \qquad (1a')$$

Algorithm: Begin by considering the module whose output corresponds to the output z of the machine. The variable representing the output signal z contains the set of states F. The input variables a_F and b_F to this module are the outputs of other modules (and/or possibly the output module) or logic constants (trivial variables) and are determined by considering Eq. (1a') and the flow table of the machine. Based on Eq. (1a'), $z \oplus a_F$ must contain the set of states which maps into F when x is 0. This implies that the variable a_F must contain the set of states that causes z to change values when x is 0. Similarly, $z \oplus b_F$ must be the set of states which maps into F when x is 1; thus b_F must contain the set of state that causes z to change values when x is 1.

The inputs to the modules whose outputs are a_F and b_F are determined by again considering Eq. (1a') and the flow table of the machine. This time the variables a_F and b_F are on the left side of the relations. One additional module is required for each non-trivial variable that appears for the first time on the right side of these relations or any relation subsequently generated. Since the output variable of each module contains a subset of the set of states of the machine and since the number of subsets of finite set is finite, the process for adding modules must eventually terminate.

Example 1. Consider the machine M_1 shown in Fig. 2a. The output module is described by the input-output relation

$$z_t = [2,3,4]_t = ([2,3,4]_{t-1} \oplus [1,3]_{t-1})\bar{x}_{t-1} + ([2,3,4]_{t-1} \oplus [\phi]_{t-1})x_{t-1}$$

This relation indicates that the inputs a_F and b_F to the output module must be the variables [1,3] and [ϕ] and this implies that there is one module in the circuit whose output signal is represented by the variable [1,3]. The variable [ϕ] is realized by the logic constant 0. The input-output relation for the module whose output variable is [1,3] is:

$$[1,3]_t = ([1,3]_{t-1} \oplus [2,4]_{t-1})\bar{x}_{t-1} + ([1,3]_{t-1} \oplus [3]_{t-1})x_{t-1}$$

Two new variables are introduced by this relation, [2,4] and [3], and therefore two more modules. Input-output relations for these modules are:

$$[2,4]_t = ([2,4]_{t-1} \oplus [2,4]_{t-1})\bar{x}_{t-1} + ([2,4]_{t-1} \oplus [3]_{t-1})x_{t-1}$$

$$[3]_t = ([3]_{t-1} \oplus [I]_{t-1})\bar{x}_{t-1} + ([3]_{t-1} \oplus [3]_{t-1})x_{t-1}$$

No additional modules are required since the right hand side of these two relations introduce no new variables; the variable [I] is realized by the logic constant 1. Figure 2b shows the AOT circuit realization of M_1. The module representing [1,3] is shaded indicating that it is initially on.

Remark. A careful examination of the above algorithm and example should reveal that when this algorithm is applied to a completely specified machine a unique AOT circuit results. Moreover, if an AOT circuit C realizes a machine M, C also realizes M_r, the reduced machine equivalent to M. This was in effect proved by Curtis[2]. In Example 1, states 2 and 4 are seen to be equivalent states.

Definition 2. An AOT module which is connected in such a way that its output never changes is called a trivial AOT module. There are two trivial AOT modules, i.e., one realizes logic constant 0 and one realizes 1.

Definition 3. An AOT circuit is called a minimum AOT circuit for a given completely specified machine if each AOT module in the realization contains a distinct non-trivial subset of the set of internal states of the machine.

Theorem 2. The minimum AOT circuit realization of a completely specified machine is unique.

3. THE TRIGGER MACHINE AND CANONICAL FORMS

Definition 4. A machine is a trigger machine if and only if the minimum AOT circuit realization of the machine is loop-free. A trigger machine is said to be of degree k if the longest chain of modules in the minimum AOT circuit is of length k.

Definition 5. An AOT tree circuit realization of a trigger machine is a loop-free AOT circuit realization in which the output of each module is an input to no more than one other module.

To obtain an AOT tree circuit from the minimum loop-free circuit one needs to replicate those modules whose outputs are inputs to more than one module.

Definition 6. By adding trivial modules an AOT tree circuit can be expanded into a tree in which all chains are of equal length. A realization of a trigger machine M in this form will be referred to as a weakly canonical AOT tree circuit realization of M. The number of rows in a weakly canonical circuit (Fig. 4) is called the depth of that circuit.

Thus for a trigger machine of degree k one can easily obtain a k-depth weakly canonical circuit realization of it. Moreover, by adding entire rows of trivial modules one can obtain a weakly canonical circuit of depth greater than k.

Definition 7. For a trigger machine M of degree k the k-depth weakly canonical AOT tree circuit realization of it is referred to as the canonical AOT circuit realization of M.

Example 2. The flow table of M_2, its minimum AOT circuit realization, an AOT tree circuit realization, and its canonical AOT tree circuit realization of depth 3 are shown in Fig. 3a, 3b, 3c, and 3d respectively.

Remark. It should be pointed out that Definitions 1-7 become definitions apropos of definite machines if the term "AOT" is replaced by "AOD" and the word "trigger" replaced by "definite". It should also be noted that there exist interesting connections between the AOD tree circuit and the work of Brzozowski[7] and Perles, Rabin and Shamir[8] on definite canonical expressions. More specifically Perles, Rabin and Shamir's canonical form corresponds to the canonical AOD circuit and Brzozowski's canonical regular expression is directly obtainable from an AOD tree circuit.

Based on the canonical AOT tree circuit realization, the canonical algebraic representation follows directly. The notation is however rather involved and thus it is worthwhile to first derive the canonical form for Example 2. We have:

for $t = 0$, $z_0 = 0$

for $t \geqslant 1$, $z_t = (z_{t-1} \oplus p_{t-1})\bar{x}_{t-1} + (z_{t-1} \oplus q_{t-1})x_{t-1}$

$$\text{for } t = 1,\ z_1 = (z_0 \oplus p_0)\bar{x}_0 + (z_0 \oplus q_0)x_0$$

$$\begin{aligned}
\text{for } t \geqslant 2,\ z_t &= (z_{t-1} \oplus p_{t-2} \oplus a_{t-2})\bar{x}_{t-1}\bar{x}_{t-2} + (z_{t-1} \oplus p_{t-2} \oplus b_{t-2})\bar{x}_{t-1}x_{t-2} \\
&\quad + (z_{t-1} \oplus q_{t-2} \oplus c_{t-2})x_{t-1}\bar{x}_{t-2} + (z_{t-1} \oplus q_{t-2} \oplus d_{t-2})x_{t-1}x_{t-2}
\end{aligned}$$

$$\begin{aligned}
\text{for } t = 2,\ z_2 &= (z_1 \oplus p_0 \oplus a_0)\bar{x}_1\bar{x}_0 + (z_1 \oplus p_0 \oplus b_0)\bar{x}_1x_0 \\
&\quad + (z_1 \oplus q_0 \oplus c_0)x_1\bar{x}_0 + (z_1 \oplus q_0 \oplus d_0)x_1x_0
\end{aligned}$$

$$\begin{aligned}
\text{for } t \geqslant 3,\ z_t &= (z_{t-1} \oplus p_{t-2} \oplus a_{t-3} \oplus 0)\bar{x}_{t-1}\bar{x}_{t-2}\bar{x}_{t-3} \\
&\quad + (z_{t-1} \oplus p_{t-2} \oplus a_{t-3} \oplus 0)\bar{x}_{t-1}\bar{x}_{t-2}x_{t-3} \\
&\quad + (z_{t-1} \oplus p_{t-2} \oplus b_{t-3} \oplus 0)\bar{x}_{t-1}x_{t-2}\bar{x}_{t-3} \\
&\quad + (z_{t-1} \oplus p_{t-2} \oplus b_{t-3} \oplus 0)\bar{x}_{t-1}x_{t-2}x_{t-3} \\
&\quad + (z_{t-1} \oplus q_{t-2} \oplus c_{t-3} \oplus 0)x_{t-1}\bar{x}_{t-2}\bar{x}_{t-3} \\
&\quad + (z_{t-1} \oplus q_{t-2} \oplus c_{t-3} \oplus 1)x_{t-1}\bar{x}_{t-2}x_{t-3} \\
&\quad + (z_{t-1} \oplus q_{t-2} \oplus d_{t-3} \oplus 0)x_{t-1}x_{t-2}\bar{x}_{t-3} \\
&\quad + (z_{t-1} \oplus q_{t-2} \oplus d_{t-3} \oplus 0)x_{t-1}x_{t-2}x_{t-3}
\end{aligned}$$

The following two definitions are presented and are then used in developing the canonical form.

Definition 8. A minterm of the variables $x_1,x_2,\cdots,x_n$ is a product term of the form $x_1^*x_2^*\cdots x_n^*$ where x_j^* is either x_j or $\bar{x}_j$. In particular, minterm $m_i(x_1,x_2,\cdots,x_n)$ is determined by writing i as an n-bit binary integer $i_1i_2\cdots i_n$. $x_j^* = x_j$ or $\bar{x}_j$ depending upon the value of i_j being 1 or 0. For example, $m_4(x_1,x_2,x_3,x_4) = \bar{x}_1x_2\bar{x}_3\bar{x}_4$.

Definition 9. Consider a weakly canonical AOT circuit of depth n shown in Fig. 4. Let the $(p + 1)^{\text{th}}$ module from the top in row r be identified by $M(r,p)$, $(p = 0,1,\cdots,2^{n-r}-1;$ and $r = 0,1,\cdots,n)$. Then $M_t(r,p)$ will denote the value of the output variable of $M(r,p)$ at time t (for $t \geqslant 0$). Note that:

(i) $M_t(0,i) = a_i$, logical constants for $t \geqslant 0$

(ii) $M_0(r,p) =$ the initial value of the module $M(r,p)$

(iii) $M_t(n,0) = z_t$, the output variable for $t \geqslant 0$

Equipped with this notation we can write an algebraic expression for a (weakly) canonical circuit of depth n (Fig. 4) as:

$$z_t = M_t(n,0) = \sum_{i=0}^{2^n-1} e_{it}m_i(x_{t-1},x_{t-2},\cdots,x_{t-n}) \qquad \text{for } t \geqslant n$$

$$\begin{aligned}
\text{where } e_{it} &= M_{t-1}(n,0) \oplus M_{t-2}(n-1,i_n) \oplus M_{t-3}(n-2,i_ni_{n-1}) \oplus \cdots \\
&\quad \oplus M_{t-n}(1,i_ni_{n-1}\cdots i_2) \oplus M_{t-n}(0,i)
\end{aligned}$$

with $i = i_ni_{n-1}\cdots i_1$, the n-bit binary integer for i.

and

$$z_k = M_k(n,0) = \sum_{j=0}^{2^k-1} e_{jk} m_j(x_{k-1}, x_{k-2}, \cdots, x_0) \quad \text{for } 0 \leq k < n$$

where $e_{jk} = M_{k-1}(n,0) \oplus M_{k-2}(n-1,j_k) \oplus M_{k-3}(n-2,j_k j_{k-1}) \oplus \cdots$

$$\oplus M_0(n-k+1, j_k j_{k-1} \cdots j_2) \oplus M_0(n-k,j)$$

with $j = j_k j_{k-1} \cdots j_1$, the k - bit binary integer for j.

Theorem 3. The canonical AOT tree circuit realization is unique for each completely specified machine, as is the canonical algebraic representation.

4. SOME PROPERTIES OF THE CLASS OF TRIGGER MACHINES

First some closure properties of the class of trigger machines are studied. Let M be a trigger machine. The degree of M is denoted as d(M).

Lemma 1. The class of trigger machines is closed under the EXCLUSIVE-OR operation. Furthermore, if M and N are trigger machines, then $d(M \oplus N) \leq \max[d(M), d(N)]$.

Proof: This theorem can be proved by direct manipulating the canonical expressions of M and N. The proof is omitted however because of its length.

Example 3. Consider two trigger machines M_3 and M_4, whose canonical circuits are shown in Fig. 5a. Each is of depth 2. We obtain the machine $(M_3 \oplus M_4)$ in Fig. 5b by forming the EXCLUSIVE-OR of the corresponding row-0 inputs and the initial values of the AOT modules of M_3 and M_4. By removing row-0 from the weakly canonical circuit in Fig. 5b, we get the canonical circuit realization of machine $(M_3 \oplus M_4)$ which is of degree 1 (Fig. 5c).

Using Lemma 1 and that $\bar{z}_t = 1 \oplus z_t$ we have the following:

Lemma 2. The class of trigger machines is closed under complementation.

Lemma 3. The class of trigger machines is closed under the operation of union. Furthermore, if M and N are trigger machines, then $d(M + N) \leq [d(M) + d(N)]$.

Proof: The proof is also omitted due to its considerable length.

Example 4. Consider two trigger machines M_5 and M_6 shown in Fig. 6a and 6b of degree of 1 and 2 respectively. The machine which is the union of M_5 and and M_6 is $M_5 + M_6$, shown in Fig. 6c and is indeed a trigger machine but note that it has degree 3 (Fig. 6d).

Theorem 4. The class of trigger machines forms a Boolean algebra.

Proof: The proof follows immediately from Lemmas 2 and 3.

Theorem 5. The class of trigger machines is neither closed under concatenation (·) nor closed under Kleene closure (*).

Proof: The reader can verify that the concatenation of machine M_5 with itself is a machine $M_5 \cdot M_5$ which is not a trigger machine. Similarly, the Kleene closure of machine M_5 is a machine M_5^* which is not a trigger machine.

Lemma 4. Let M be a reduced machine with the set of internal states $S = [s_1, s_2, \cdots, s_n]$. If M is a trigger machine, then (1) a machine M_i with the same set of internal states as M but with a single final state $[s_i]$ (for any i) will also be a trigger machine; (2) a machine M_j with S as the set of internal states but with any subset of S as the set of final states is also a trigger machine.

Proof: To prove (1) let $A_1, A_2, \cdots, A_k$ be the set of AOT modules such that a loop-free connection of them realizes M. Any machine using A_i ($1 \leq i \leq k$) as the output module (also denoted as machine A_i) should also be a trigger machine by definition. When M is in state s_i the outputs of those modules $A_{i_1}, A_{i_2}, \cdots, A_{i_j}$ containing s_i will be on and others will be off. We can express the output of machine M_i with a single final state s_i in terms of A_i's (note that this may not be true when the machine is not reduced), i.e., $M_i = A_1^* A_2^* \cdots A_k^*$ where $A_p^* = A_p$ if $p \varepsilon [i_1, i_2, \cdots, i_j]$; and $A_p^* = \bar{A}_p$ otherwise. M_i is a Boolean function of trigger machines $A_1, A_2, \cdots, A_k$, hence a trigger machine by Theorem 4.

Part (2) of the Lemma is a direct extension of part (1).

Theorem 6. A necessary condition for a reduced machine to be a trigger machine is that every next state column in the flow table must be a permutation column.

Proof: To prove this we first assume that the converse is true, namely, that a reduced machine M with a non-permutation column can still be a trigger machine. Let M be such a trigger machine having the set of states $S = [s_1, s_2, \cdots, s_n]$. With no loss of generality let the next state column under $x = 0$ be a non-permutation column. In particular, assume states $s_{i_1}, s_{i_2}, \cdots, s_{i_j}$ do not appear as the next state entries in column $x = 0$. By Lemma 4 the machine M_i with the same state behavior as M and with $F_i = [s_{i_1}, s_{i_2}, \cdots, s_{i_j}]$ as the set of final states is a trigger machine. Since no states map into F_i when $x=0$ it follows that $z_i \oplus a_{F_i} = [\phi]$ and thus $a_{F_i} = z_i$, which in turn implies that a loop exists in the AOT realization of M_i. This contradicts the assumption that M is a trigger machine.

As a result of this theorem and the fact that a definite machine can not have a permutation column we obtain the following consequence:

Corollary 1. The class of trigger machines and the class of definite machines are disjoint.

This fact was presented by Harlow and Coates (Result 36 of [1]) although a different approach is used in their proof.

5. A TESTING PROCEDURE

Theorem 7. If a trigger machine is of degree k, then it has at least k + 1 internal states.

Proof: Let M be an n state trigger machine. Let $(s_1, s_2, \cdots, s_n)$ be a fixed ordering of the set of internal states of M. Let $v(M_a) = (v_1, v_2, \cdots, v_n)$, corresponding to a module M_a in the AOT realization of M, be an n-dimensional vector over GF(2). The component $v_i = 1$ if the output of module M_a contains state s_i; $v_i = 0$ otherwise. The mod-2 sum of two vectors $v(M_a)$ and $v(M_b)$ is the vector corresponding to $M_a \oplus M_b$, or $v(M_a) \oplus v(M_b) = v(M_a \oplus M_b)$. Assume the degree of M is k. Thus there exist at least one path in the AOT realization of M with exactly k non-trivial modules. Let us choose one such path and designate those modules in the chain from left to right as $M_{p1}, M_{p2}, \cdots, M_{pk}$. Note that at least one of the two inputs to M_{p1} must be a logic constant 1 which is denoted as M_I. $v(M_I) = (1, 1, \cdots, 1)$. Now the following statement is true: $v(M_I), v(M_{p1}), \cdots, v(M_{pk})$ are linearly independent vectors. If not, then some vector $v(M_{pi})$ can be expressed as linear combination of the other vectors:

$$v(M_{pi}) = C_I v(M_I) \oplus C_1 v(M_{p1}) \oplus \cdots \oplus C_{i-1} v(M_{pi-1})$$

where C_I and C_j's are constants and not all of them are zero. This expression is equivalent to

$$M_{pi} = C_I M_I \oplus C_1 M_p \oplus \cdots \oplus C_{i-1} M_{pi-1}$$

which is impossible because, by Lemma 1, M_{pi} would be realizable by a shorter chain. Therefore $v(M_I), v(M_{p1}) \cdots, v(M_{pk})$ are $k + 1$ linearly independent vectors. It then follows that the dimension of the vectors must be at least $k + 1$ or $n \geqslant k + 1$.

The corresponding theorem for definite machines was proved using three different methods in Theorem 1 of [8]. A fourth method similar to the above is thus possible.

As the direct consequence of this theorem we have the following result:

Corollary 2. Given an n state machine M, an algorithm requiring at most $2^{n-1} - 1$ steps exists which can determine whether M is a trigger machine.

Proof: By Theorem 7 if M is a trigger machine with n states then M can not have more than n-1 non-trivial rows in the AOT tree realization. Thus for a given n state machine one need only apply the realization algorithm for at most n-1 rows. If a loop occurs among the first n-1 rows of modules or the realization does not terminate after n-1 rows, then the machine is not a trigger machine. A tree of depth n-1 has at most $2^{n-1} - 1$ modules and the connection of each module can be determined by a single simple test.

Using Harlow and Coates' notion (Result 33 of [1]) one is also able to test whether a given machine belongs to the class of trigger machines. However, our procedure is more efficient.

6. AN ALTERNATE DEFINITION FOR THE TRIGGER MACHINE

Definition 10. A trigger flip-flop_(T-FF) is a two state machine with a single input a, two outputs, b and b, having the following input-output relation:

$$b_t = b_{t-1} \oplus a_{t-1} \qquad \text{for } t \geqslant 1$$

$$b_0 = 0 \text{ or } 1$$

Theorem 8. A synchronous sequential machine is a trigger machine if and only if there exist a loop-free circuit of trigger flip-flops and combinational gates that can realize it.

Proof: To prove this theorem it is sufficient to prove that any loop-free circuit of T-FF's and gates can be converted into a loop-free connection of AOT modules. It is necessary to consider only the following possible cases. (Note also that only the positive output of a T-FF, b_t, is considered in the following because that if b_t represents a trigger variable so does $\bar{b}_t$ by Lemma 2.)

(a) If the input to a T-FF is a trigger variable a (including the constants 0 and 1), then we can express the output variable b as:

$$b_t = b_{t-1} \oplus (a_{t-1}\bar{x}_{t-1} + a_{t-1}x_{t-1})$$

This can be directly realized by an AOT module as shown in Fig.7a.

(b) If the input to the T-FF is the external input x (or $\bar{x}$) then we write:

$$b_t = b_{t-1} \oplus x_{t-1} = b_{t-1} \oplus (0 \cdot \bar{x}_{t-1} + 1 \cdot x_{t-1})$$

or

$$b_t = b_{t-1} \oplus \bar{x}_{t-1} = b_{t-1} \oplus (1 \cdot \bar{x}_{t-1} + 0 \cdot x_{t-1})$$

Fig.7b shows these two cases and their corresponding AOT circuits.

(c) When the inputs to a combinational circuit are trigger variables a,b,c,$\cdots$,w (Fig. 7c), then by Theorem 4, the output of the combinational circuit also represents a trigger machine.

(d) The last possible case occurs when the inputs to a combinational circuit which is followed by a T-FF are trigger variables and the external input x. In this case, let the combinational logic be designated by the function f. The output of the combinational logic at time t is thus (Fig. 7d)

$$f(a_t, b_t, \cdots, w_t, x_t)$$

The output of the T-FF will be

$$g_t = g_{t-1} \oplus f(a_{t-1}, b_{t-1}, \cdots, w_{t-1}, x_{t-1})$$

By expanding about the input x we write

$$g_t = g_{t-1} \oplus [f(a_{t-1}, b_{t-1}, \cdots, w_{t-1}, 0)\bar{x}_{t-1} + f(a_{t-1}, b_{t-1}, \cdots, w_{t-1}, 1)x_{t-1}]$$

This equation can be realized by an AOT module if $a = f(a_{t-1}, b_{t-1}, \cdots, w_{t-1}, 0)$ and $b = f(a_{t-1}, b_{t-1}, \cdots, w_{t-1}, 1)$.

Example 5. Fig. 8 illustrates the conversion of a circuit consisting of three T-FF's and few logic gates to a canonical AOT circuit.

If terms are appropriately changed in Theorem 8, then it reads as "A synchronous sequential machine is a definite machine if and only if there exists a loop-free circuit of unit delays and combinational gates that can realize it." This is a well-known definition of the definite machine.

REFERENCES

[1] Harlow, C. A. and Coates, C. L., "On the Structure of Sequential Machine Realizations," Laboratories for Electronics and Related Science Research, Univ. of Texas, Austin, Texas, Tech. Report 27, July 1967.

[2] Curtis, H. A., "The Realization of Polylinear Sequential Circuits Using Flip-Flop Memory," IEEE TEC. Jan. 1971, pp. 87-94.

[3] Ullman, J. D., and Weiner, P., "Universal Two-State Machines: Characterization Theorems and Decomposition Schemes," 9th Annual IEEE Symposium on Switching and Automata Theory, Schenectady, New York, Oct. 1968, pp. 413-426.

[4] Newborn, M. M., "A Synthesis Technique for Binary Input-Binary Output Synchronous Sequential Moore Machines," IEEE TEC-17, No. 7, July 1968, pp. 697-699.

[5] Weiner, P. and Hopcroft, J. E., "Modular Decomposition of Synchronous Sequential Machines," 8th Annual IEEE Symp. on Switching and Automata Theory, Austin, Texas, Oct. 1967, pp. 233-239.

[6] Arnold, T. F., Tan, C. J., and Newborn, M. M., "Iteratively Realized Sequential Circuits," 9th Annual IEEE Symposium on Switching and Automata Theory, Schenectady, New York, Oct. 1968, pp. 431-448.

[7] Brzozowski, J. A., "Canonical Regular Expressions and Minimal State Graphs for Definite Events," Proc. Symp. on Mathematical Theory of Automata, Microwave Research Inst. Symp. Ser., vol. 12. Brooklyn, N.Y., Polytechnic Press, 1962, pp. 529-561.

[8] Perles, M., Rabin, M. O., and Shamir, E., "The Theory of Definite Automata," IEEE TEC, June 1963, pp. 233-243.

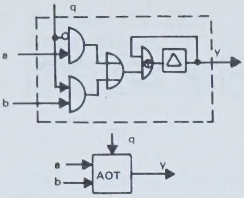
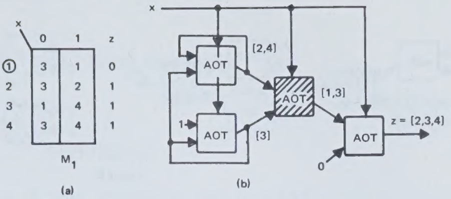

q
a
b
y
Δ
q
a
b
AOT
y

Figure 1. An AND-OR-TRIGGER module



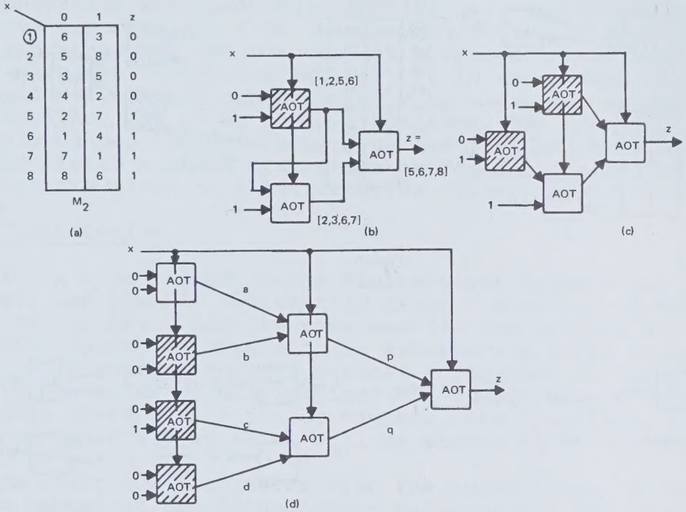
x	0	1	z
①	3	1	0
2	3	2	1
3	1	4	1
4	3	4	1

M_1

(a)

x
AOT
[2,4]
AOT
[1,3]
1
AOT
[3]
AOT
z = [2,3,4]
0
(b)

Figure 2. (a) Machine M_1 and (b) an AOT realization



x	0	1	z
①	6	3	0
2	5	8	0
3	3	5	0
4	4	2	0
5	2	7	1
6	1	4	1
7	7	1	1
8	8	6	1

M_2

(a)

x
[1,2,5,6]
0
1
AOT
AOT
z =
[5,6,7,8]
1
AOT
[2,3,6,7]
(b)
x
0
1
AOT
AOT
z
0
1
AOT
AOT
1
(c)
x
0
0
AOT
a
AOT
0
0
AOT
b
p
AOT
z
0
1
AOT
c
AOT
q
0
0
AOT
d
(d)

Figure 3. (a) Machine M_2, (b) the minimum AOT realization of M_2, (c) an AOT tree circuit realization, and (d) the canonical AOT circuit realization

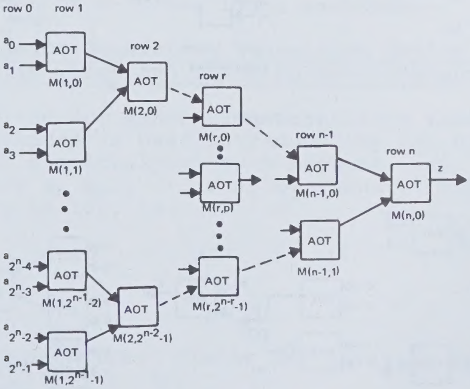

row 0
row 1
a_0
a_1
AOT
M(1,0)
row 2
AOT
M(2,0)
row r
AOT
M(r,0)
a_2
a_3
AOT
M(1,1)
row n-1
AOT
M(n-1,0)
row n
AOT
M(n,0)
z
AOT
M(r,p)
AOT
M(n-1,1)
a_{2^n-4}
a_{2^n-3}
AOT
$M(1,2^{n-1}-2)$
AOT
$M(2,2^{n-2}-1)$
AOT
$M(r,2^{n-r}-1)$
a_{2^n-2}
a_{2^n-1}
AOT
$M(1,2^{n-1}-1)$

Figure 4. A (weakly) canonical AOT circuit of depth n.
(The input x is an implicit input to all modules.)

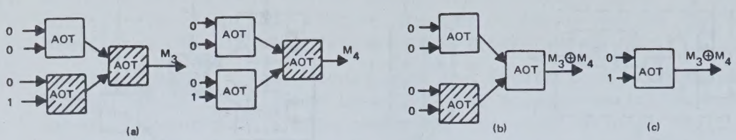
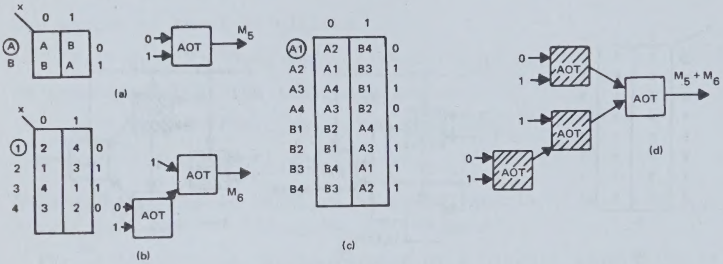

0
0
AOT
0
1
AOT
AOT
M_3
(a)
0
0
AOT
0
1
AOT
AOT
M_4
0
0
AOT
0
0
AOT
AOT
$M_3 \oplus M_4$
(b)
0
1
AOT
$M_3 \oplus M_4$
(c)

Figure 5


x
0
1
AOT
M_5
(a)
AOT
AOT
M_6
(b)
(c)
AOT
AOT
AOT
AOT
$M_5 + M_6$
(d)

x	0	1	
Ⓐ	A	B	0
B	B	A	1

x	0	1	
①	2	4	0
2	1	3	1
3	4	1	1
4	3	2	0

	0	1	
(A1)	A2	B4	0
A2	A1	B3	1
A3	A4	B1	1
A4	A3	B2	0
B1	B2	A4	1
B2	B1	A3	1
B3	B4	A1	1
B4	B3	A2	1

Figure 6

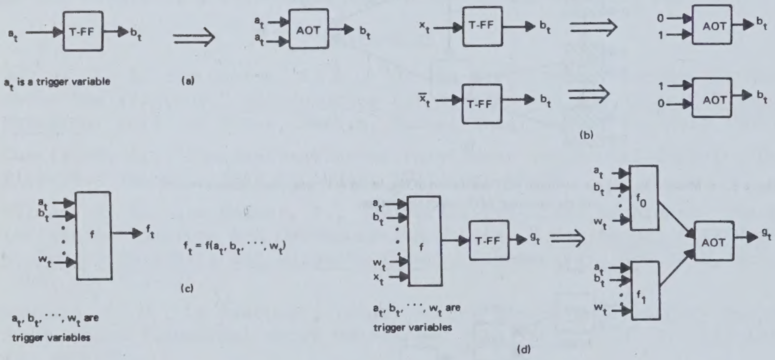

a_t
T-FF
b_t
a_t is a trigger variable
(a)
a_t
a_t
AOT
b_t
x_t
T-FF
b_t
0
1
AOT
b_t
$\bar{x}_t$
T-FF
b_t
1
0
AOT
b_t
(b)
a_t
b_t
w_t
f
f_t
$f_t = f(a_t, b_t, \cdots, w_t)$
(c)
$a_t, b_t, \cdots, w_t$ are trigger variables
a_t
b_t
w_t
x_t
f
T-FF
g_t
f_0
f_1
AOT
g_t
$a_t, b_t, \cdots, w_t$ are trigger variables
(d)

Figure 7

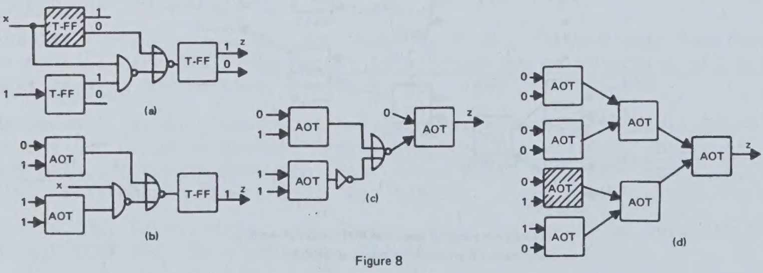

x
T-FF
1
0
T-FF
1
z
0
1
T-FF
1
0
(a)
0
1
AOT
1
1
AOT
x
T-FF
1
z
(b)
0
1
AOT
1
1
AOT
0
AOT
z
(c)
0
0
AOT
0
0
AOT
AOT
0
1
AOT
1
0
AOT
AOT
AOT
z
(d)

Figure 8

ALGORITHMS FOR CONSTRUCTING TRANSITION TABLES OF MINIMUM-STATE COMPUTING SEQUENTIAL MACHINES

SAMUEL C. LEE
Department of Electrical Engineering
University of Houston
Houston, Texas 77004

ABSTRACT

This paper describes some simple algorithms for constructing transition tables of minimum-state two-valued and m-valued logic multiplying- and dividing-by-N machines. The algorithms for generating m-valued logic multiplying- and dividing-by-N machines are obtained from the generalization of the patterns of the next-states and the outputs observed in the two-valued logic machines transition tables. It is shown that one needs at least N states in constructing a multiplying- or dividing-by-N machine. By using these algorithms, the transition tables of any m-valued logic multiplying- and dividing-by-N machine for any finite m and N transition tables so obtained is N. Examples are given to illustrate the algorithms.

I. Introduction

It is well known* that a finite-state machine can not multiply two numbers arbitrarily large. However, if the multiplier is finite, a finite-state machine can perform any multiplication operation even with an arbitrarily large multiplicand. This paper describes several algorithms for generating transition tables of multiplying- and dividing-by-N sequential machines, where N is a finite number. The transition tables obtained use N states which will be proven to be minimum.

We start our discussion with the construction of transition tables of two-valued logic multiplying- and dividing-by-N machines. Two simple algorithms for generating transition tables using minimum number of states are presented. From the patterns of the next-states and outputs of these transition tables, the algorithms are generalized to the construction of transition tables of m-valued logic machines.

II. Algorithms for Generating Transition Tables of Two-Valued Logic Multiplying- and Dividing-by-N Sequential Machines

The algorithm for generating transition tables of multiplying-by-N machines is best introduced by use of an example. Let us consider a multiplying-by-N machine, say, N = 5. Any multiplicand $a = a_n a_{n-1} \ldots a_2 a_1$ (binary) multiplied by 5, which in binary is 101, is

---

*M. Minsley, Computation: Finite and Infinite Machines. Prentice-Hall, 1967, p. 26.

$$\begin{array}{lcccccccc} & a_n & a_{n-1} & a_{n-2}\cdots a_{i+2} & a_{i+1}a_i & \cdots & a_3 & a_2 & a_1 \\ \times & & & & & & 1 & 0 & 1 \\ \hline & a_n & a_{n-1} & a_{n-2}\cdots a_{i+2} & a_{i+1}a_i & \cdots & a_3 & a_2 & a_1 \end{array}$$

$$\begin{array}{cccccccccc} a_n & a_{n-1} & a_{n-2} & \cdots\cdots & a_i & \cdots & a_3 & a_2 & a_1 & \\ \hline a_n & a_{n-1} & a_n+a_{n-2} & \cdots\cdots & a_{i+2}+a_i & \cdots & a_3+a_1 & a_2 & a_1 \end{array}$$

Except the first two and the last two digits which are essentially copying from the multiplicant $a_n \ldots a_1$. All the rest of the terms in between are resulted from the sums of $a_{i+2}+a_i$, $i = 1, 2, \ldots, n-2$. Let us consider the term $a_3 + a_1$ which is the output at the time a_3 is applied. For a_3 is 0 and 1, the outputs are a_1 and $1+a_1$, respectively. In general, for any input a_i, $i = 3, 4, \ldots, n$, the corresponding output is a_{i-2} if a_i is 0, and is $1 + a_{i-2}$ if a_i is 1.

Now the question is what are the states of this machine and what is the next-state function? To answer the first question, we find that we need five states represented by 00, 01, 10, 11, and 100. The output and next-state upon an application of an input a_i may be obtained as follows.

1. If a_i is 0, the output and the next-state are obtained by

   Input the 0 here ↓
   $$\underbrace{0\ X}_{\text{next state}}\ \underset{\text{output}}{\boxed{X}} \qquad \text{Input the 0 here} \downarrow \quad \underbrace{0\ 1\ 0}_{\text{next state}}\ \underset{\text{output}}{\boxed{0}}$$

   where X X denotes the present state which may be any one of the four states 00, 01, 10, and 11.

2. If a_i is 1, the output and the next state are then obtained by

   $$\begin{array}{rccc} & X & X & \\ + & 1 & 0 & 1 \\ \hline & \multicolumn{2}{c}{\underbrace{c_3\ \ c_2}_{\text{next state}}} & \underset{\text{output}}{\boxed{c_1}} \end{array} \qquad \text{or} \qquad \begin{array}{cccc} & 1 & 0 & 0 \\ & 1 & 0 & 1 \\ \hline \multicolumn{3}{c}{\underbrace{1\ 0\ 0}_{\text{next state}}} & \underset{\text{output}}{\boxed{1}} \end{array}$$

   Constructing in this way, the complete state transition table can be obtained and is shown in Table I(a).

By generalizing the above procedure, we have the following simple algorithm.

Algorithm 1. In constructing multiplying-by-N machine, N states are used which are represented in binary by

$$\left.\begin{array}{c} 0\ 0\ 0\ .\ .\ .\ 0\ 0 \\ 0\ 0\ 0\ .\ .\ .\ 0\ 1 \\ 0\ 0\ 0\ .\ .\ .\ 1\ 0 \\ \hline \text{up to } (N-1) \end{array}\right\} \text{N numbers}$$

which may be expressed as

$$q_i = y_k^{\,i} \cdots y_2^{\,i}\, y_1^{\,i}$$

where $y_j^{\,i}$ is 0 or 1. The input a is represented in binary as

$$a = x_n \ . \ . \ . \ x_2\, x_1.$$

When $x_i = 0$

Step 1A: Place x_i on the left side of the $y_k^{\,i}$.

Step 2A: Output $z_i = y_1^{\,i}$, the least significant bit of this concentration.

Step 3A: Next state $q_{i+1} = x_i\, y_k^{\,i}\, y_{k-1}^{\,i} \ . \ . \ . \ y_3^{\,i}\, y_2^{\,i}$

When $x_i = 1$

Step 1B: Perform binary addition on q_i and N.

Step 2B: Let the result be $r_{k+1}\, r_k\, r_{k-1} \ . \ . \ . \ r_3 r_2 r_1$ (Note that the maximum number of bits will be (k+1))

Step 3B: Output z_i = least significant bit of the addition (i.e., r_1)

Step 4B: Next state $q_{i+1} = r_k\, r_{k-1} \ . \ . \ . \ r_2 \ . \ . \ . \ r_1$

A pictorial representation of this algorithm is given in Fig. 1

In a similar manner, we can derive the algorithm for constructing the transition tables for dividing machines.

Algorithm 2. The procedure for constructing dividing-by-N machines using N states is as follows:

Step 1: Place x_i on the right side of the $y_1^{\,i}$ bit thereby shifting all $y_j^{\,i}$ to the left side

Step 2: Output z_i = the most significant bit of this shift (i.e., $z_i = y_k^{\,i}$)

If $z_i = y_k^{\,i} = 0$,

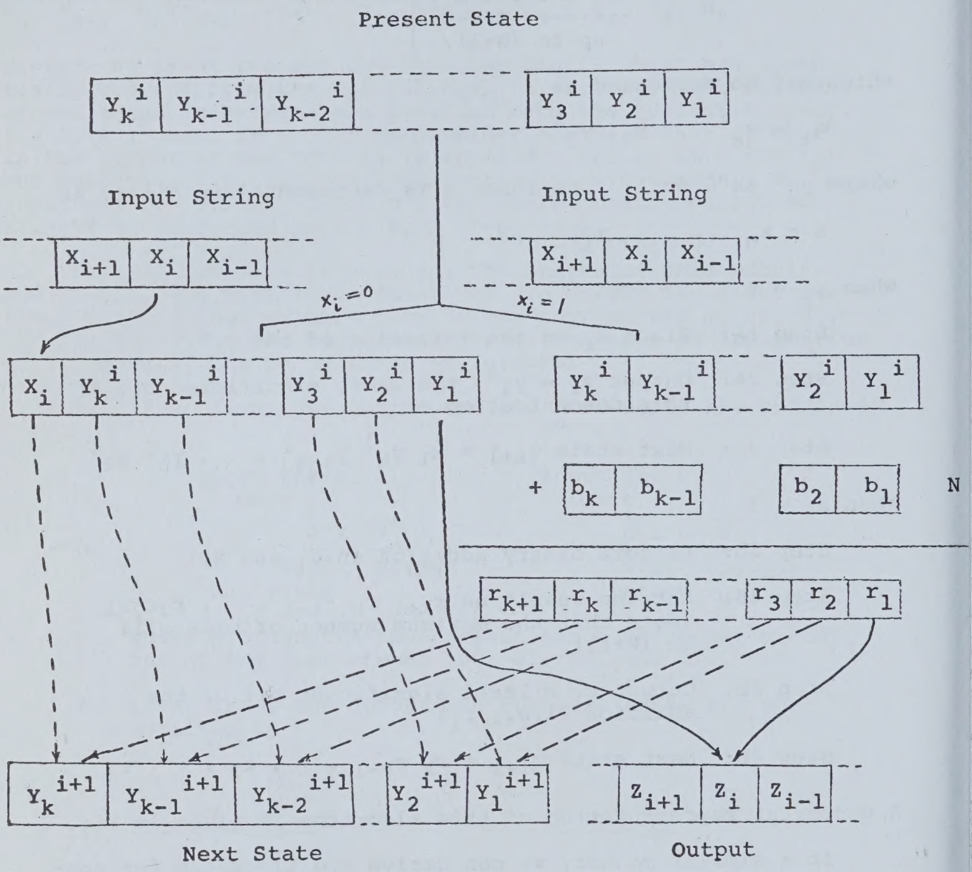

Present State
Y_k^i Y_{k-1}^i Y_{k-2}^i Y_3^i Y_2^i Y_1^i
Input String
X_{i+1} X_i X_{i-1}
Input String
X_{i+1} X_i X_{i-1}
$x_i = 0$
$x_i = 1$
X_i Y_k^i Y_{k-1}^i Y_3^i Y_2^i Y_1^i
Y_k^i Y_{k-1}^i Y_2^i Y_1^i
+ b_k b_{k-1} b_2 b_1
N
r_{k+1} r_k r_{k-1} r_3 r_2 r_1
Y_k^{i+1} Y_{k-1}^{i+1} Y_{k-2}^{i+1} Y_2^{i+1} Y_1^{i+1}
Z_{i+1} Z_i Z_{i-1}
Next State
Output

Fig. 1. A Pictorial Representation of Algorithm 1

Step 3A: Next state $q_{i+1} = y_{k-1}^{i} \, y_{k-2}^{i} \, . \, . \, . \, y_1^{i} \, x_1$

If $z_i = y_k^{i} = 1$,

Step 3B: From $y_k^{i} \, y_{k-1}^{i} \, . \, . \, . \, y_2^{i} \, y_1^{i} \, x_i$, subtract $b_1 \, b_2 \, . \, . \, . \, b_k$.

Step 4B: Let the result be $r_k \, r_{k-1} \, r_{k-2} \, . \, . \, . \, r_1$. (It will always be of length $\leq k$, in this case.)

Step 5B: Next state $q_{i+1} = r_k \, r_{k-1} \, . \, . \, . \, r_2 \, r_1$.

A pictorial representation of this algorithm is shown in Fig. 2.

III. Patterns of Next-States and Outputs of the Transition Tables Obtained by Algorithms 1 and 2

Some interesting patterns of the next-states and outputs of the transition tables obtained by Algorithms 1 and 2 should be brought out to our special attention. For example, the transition tables of multiplying-by-5, 6, and 7 machines and those of dividing machines dividing by the same numbers obtained by Algorithms 1 and 2 are shown in Table I. The following interesting patterns of the next states and outputs are observed.

1. Patterns of Transition Table of Binary Multiplying-by-N Machines

   A. The next states in the transition table have the following pattern.

Present State	0 Input	1 Input
00	00	$\begin{cases} \frac{N-1}{2} & \text{odd } N \\ \frac{N-2}{2} & \text{even } N \end{cases}$
01	00	.
10	01	.
11	01	.
.	.	.
.	.	.
.	.	.
N-1	$\begin{cases} \frac{N-1}{2} & \text{odd } N \\ \frac{N-2}{2} & \text{even } N \end{cases}$	N-1

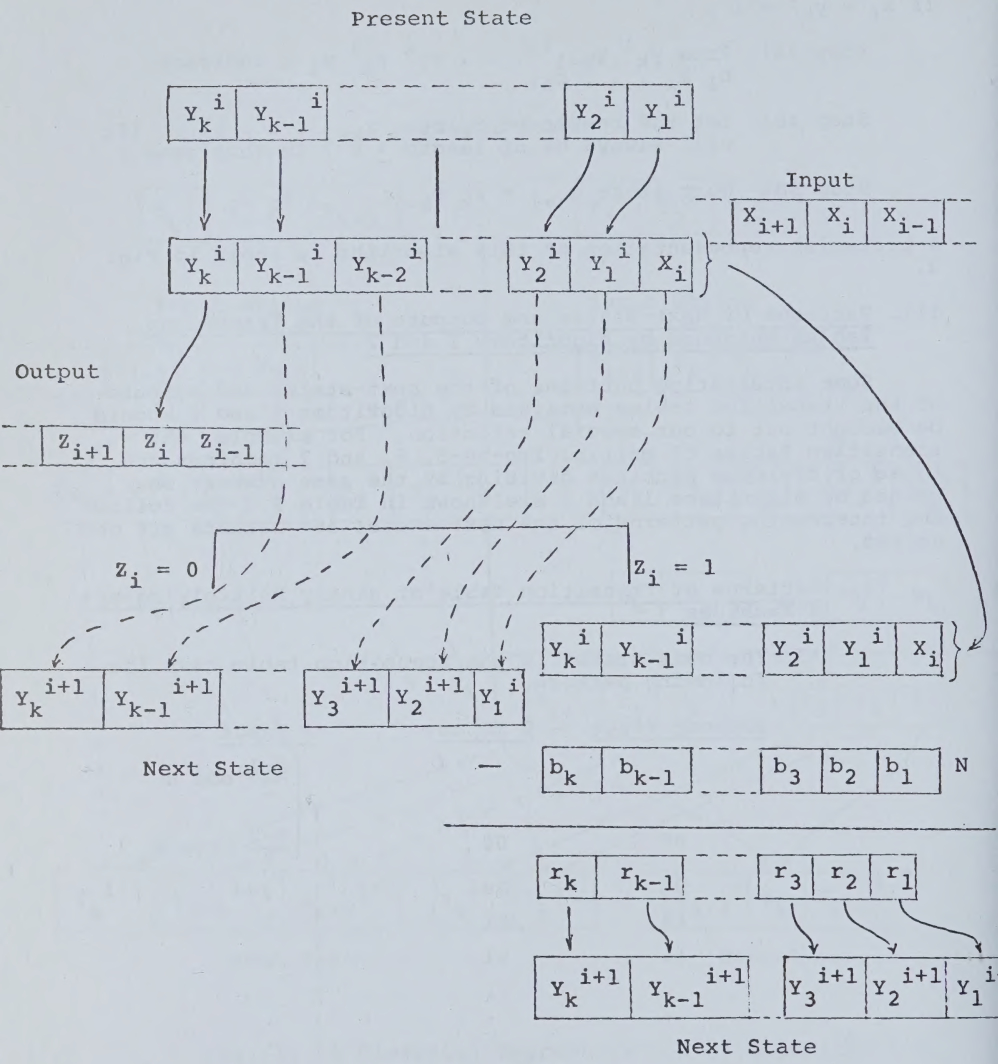

Present State
Y_k^i
Y_{k-1}^i
Y_2^i
Y_1^i
Input
X_{i+1}
X_i
X_{i-1}
Y_k^i
Y_{k-1}^i
Y_{k-2}^i
Y_2^i
Y_1^i
X_i
Output
Z_{i+1}
Z_i
Z_{i-1}
$Z_i = 0$
$Z_i = 1$
Y_k^i
Y_{k-1}^i
Y_2^i
Y_1^i
X_i
Y_k^{i+1}
Y_{k-1}^{i+1}
Y_3^{i+1}
Y_2^{i+1}
Y_1^i
Next State
—
b_k
b_{k-1}
b_3
b_2
b_1
N
r_k
r_{k-1}
r_3
r_2
r_1
Y_k^{i+1}
Y_{k-1}^{i+1}
Y_3^{i+1}
Y_2^{i+1}
Y_1^i
Next State

Fig. 2. A Pictorial Representation of Algorithm 2

B. The outputs in the transition table have the pattern that 0's and 1's are alternating, starting with 0 and ending with 1.

From the transition tables of dividing-by-N machines shown in the right column of Table I, we discover the following patterns.

2. Pattern of Transition Table of Binary Dividing-By-N Machines

A. The next states in the transition table have the following pattern. For Even N

Present State	0 Input	1 Input
00	00	01
01	10	11
100	100	101
	even number	odd number
⋮	⋮	⋮
$\frac{N}{2} - 1$	N-2	N-1
	The Same	The Same
$\frac{N}{2}$	00	01
$\frac{N}{2} + 1$	10	11
	Even Number	Odd Number
$\frac{N}{2} + 2$	100	101
⋮	⋮	⋮
N-1	N-2	N-1

For Odd N

Present State	0 Input	1 Input
00	00	01
01	10	11
10	100	101
	even number	odd number
⋮	⋮	⋮
$\frac{N-3}{2}$	N-3	N-2
$\frac{N-1}{2}$	N-1	00
$\frac{N+1}{2}$	01	10
	odd number	even number
$\frac{N+3}{2}$	11	100
⋮	⋮	⋮
N-1	N-2	N-3

TABLE I

RANSITION TABLES OF SEVERAL MULTIPLYING-BY-X AND DIVIDING-BY-N MACHINES

x_k q_k	0	1	0	1
00	00	10	0	1
01	00	11	1	0
10	01	11	0	1
11	01	100	1	0
100	10	100	0	1

A. Multiplying-By-5 Machine (N = 101)

x_k q_k	0	1	0	1
00	00	01	0	0
01	10	11	0	0
10	100	00	0	1
11	01	10	1	1
100	11	100	1	1

B. Dividing-By-5 Machine (N = 101)

x_k q_k	0	1	0	1
00	00	11	0	0
01	00	11	1	1
10	01	100	0	0
11	01	100	1	1
100	10	101	0	0
101	10	101	1	1

C. Multiplying-By-6 Machine (N = 110)

x_k q_k	0	1	0	1
00	00	01	0	0
01	10	11	0	0
10	100	101	0	0
11	00	01	1	1
100	10	11	1	1
101	100	101	1	1

D. Dividing-By-6 Machine (N = 110)

x_k q_k	0	1	0	1
00	00	11	0	1
01	00	100	1	0
10	01	100	0	1
11	01	101	1	0
100	10	101	0	1
101	10	110	1	0
110	11	110	0	1

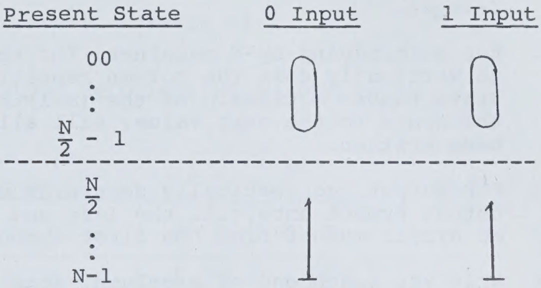
E. Multiplying-By-7 Machine (N = 111)

x_k q_k	0	1	0	1
00	00	01	0	0
01	10	11	0	0
10	100	101	0	0
11	110	00	0	1
100	01	10	1	1
101	11	100	1	1
110	101	110	1	1

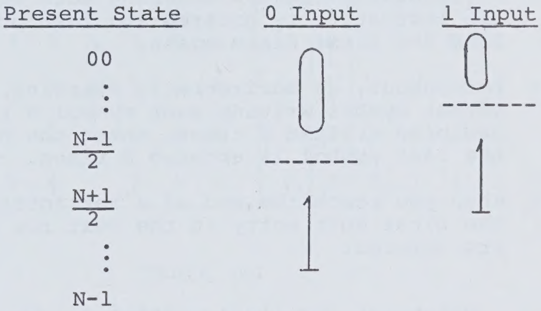
F. Dividing-By-7 Machine (N = 111)

B. The outputs in the transition table have the pattern shown below.

For Even N


Present State
0 Input
1 Input
00
$\frac{N}{2} - 1$
$\frac{N}{2}$
N-1

For Odd N


Present State
0 Input
1 Input
00
$\frac{N-1}{2}$
$\frac{N+1}{2}$
N-1

These patterns not only enable us to write down the transition tables of a multiplying- and dividing-by-N machine with arbitrary (finite) N directly, but more importantly give us a clue for constructing transition tables of m-valued logic machines which is discussed in the next section.

IV. M-Valued Logic Computing Sequential Machines

In the above, the computing machines discussed were binary machines. The extension of the construction of binary multiplying-by-N and dividing-by-N machines to m-valued logic is a nontrivial problem. However, with the aid of the patterns of the next-states and outputs in the binary multiplying- and dividing-by-N machines transition tables, the following rules can be obtained:

$$Q = \{0, 1, 2, \ldots, N-1\}$$

and

$$X = Z = \{0, 1, 2, \ldots, m-1\}$$

Algorithm 3. The rules for constructing transition tables of m-valued logic multiplying- and dividing-by-N sequential machines is as follows:

Rule 1: Start at the top left most corner showing 0-state, 0-input.

Rule M_2: For multiplying-by-N machines, for the next states, go vertically down the column repeating the same state number m times. At the (m+1)th entry, change the state to the next value; till all the states have been written.

Rule M_3: For output, go vertically downwards entering each output symbol once till the last one and then proceed in cyclic manner from the first symbol again.

Rule M_4: When you reach end of a column, start at the top of the next column till all columns have been entered.

Rule D_2: For dividing-by-N machines, for next states, go horizontally forward entering each state once, till the last state is entered and then proceed cyclically from the first state again.

Rule D_3: For output, go horizontally starting from the first output symbol writing each symbol N times; when it has been written N times, enter the next symbol, till the last symbol is entered N times.

Rule D_4: When you reach the end of a horizontal row, start at the first left entry in the next row till all blocks are entered.

Example 1. Construct a decimal multiplying-by-7 machine. The (minimal) number of states required is 7; they are denoted by 0, 1, 2, . . . , 6 states, or

$$Q = \{0, 1, 2, . . . , 6\}$$

The input and output spaces are

$$X = Z = \{0, 1, 2, . . . , 9\}$$

By applying the rules stated above, the transition table of this machine is shown in Table II.

Example 2. As another example, consider a decimal dividing-by-6 machine. The number of states required in this case is 6, they are

$$Q = \{0, 1, 2, . . . , 5\}$$

The input and output spaces are

$$X = Z = \{0, 1, 2, . . . , 9\}$$

TABLE II

DECIMAL MULTIPLYING-BY-7 MACHINE

	Next State										Output									
q_k \ x_k	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
0	0	0	1	2	2	3	4	4	5	6	0	7	4	1	8	5	2	9	6	3
1	0	0	1	2	2	3	4	5	5	6	1	8	5	2	9	6	3	0	7	4
2	0	0	1	2	3	3	4	5	5	6	2	9	6	3	0	7	4	1	8	5
3	0	1	1	2	3	3	4	5	5	6	3	0	7	4	1	8	5	2	9	6
4	0	1	1	2	3	3	4	5	6	6	4	1	8	5	2	9	6	3	0	7
5	0	1	1	2	3	4	4	5	6	6	5	2	9	6	3	0	7	4	1	8
6	0	1	2	2	3	4	4	5	6	6	6	3	0	7	4	1	8	5	2	9

TABLE III

DECIMAL DIVIDING-BY-7 MACHINE

q_k \ x_k	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	0	1	2	3	0	0	0	0	0	0	1	1	1	1
1	4	5	0	1	2	3	4	5	0	1	1	1	2	2	2	2	2	2	3	3
2	2	3	4	5	0	1	2	3	4	5	3	3	3	3	4	4	4	4	4	4
3	0	1	2	3	4	5	0	1	2	3	5	5	5	5	5	5	6	6	6	6
4	4	5	0	1	2	3	4	5	0	1	6	6	7	7	7	7	7	7	8	8
5	2	3	4	5	0	1	2	3	4	5	8	8	8	8	9	9	9	9	9	9

The transition table is shown in Table III.

Theorem. The minimal number of states required for constructing a m-valued logic multiplying- and dividing-by-N machine is N.

This theorem can be proven by several ways. The simplest, yet it is still a vigorous one, is to show that the N states in the multiplying-by-N or dividing-by-N tables obtained by the algorithms presented in this paper are distinguishable. In other words, none of them are equivalent. This can be easily seen from the patterns of next-states and outputs of the transition tables presented above.

V. Conclusion

Several algorithms for constructing transition tables of z-valued and m-valued logic computing sequential machines have been presented. By using them, the transition tables of any multiplying- and dividing-by-N machines can be obtained using N states. The machines so obtained are minimal, since all of its states are distinguishable. Examples have been given to illustrate the algorithms.

A HEURISTIC APPROACH TO MULTIMODAL FUNCTION MINIMIZATION- THE GLOBAL ASPECT

JASNA OPAČIĆ
Electrical Engineering Department
University of Maryland
College Park, Maryland 20742

Abstract

This paper presents a heuristic approach to a search of a multimodal surface. Such a kind of surface presents non-trivial difficulties which become increasingly difficult to handle as the number of dimensions increases. When the aim is the global extremum, a majority of available techniques use a probabilistic approach, which is doomed by "the curse of dimensionality." The proposed strategy is not affected by the problem of dimensionality, at least no more than is a second-order gradient descent technique, which is used for unimodal minimization. The proposed strategy handles the global minimization through several heuristic postulates which guide a search in an adaptive manner.

INTRODUCTION

For nonlinear optimization problems with multimodel performance index structure analytical solutions are not readily available and some kind of a global search strategy is required. The problem has already been tackled by a random approach [1], [2], by a grid-type probabilistic approach [3], [4] and one recent extension of [4] involves an adaptive scheme with learning features [5]. However, the efficiency of the available methods of which the author is aware, is critically dependent on the dimensionality of the parameter space, with the consequences that even for a small number of parameters a prohibitive computer time, or storage, or both, are usually required.

Engineering applications often call for more practical schemes which while being perhaps somewhat less rigorous more than compensate for this by much greater efficiency. The scheme which is proposed is based on a repetitive implementation of heuristic assumptions about a given surface and subsequent evaluations of results obtained. It appears that the heurietic method of problem handling [6], that permits an investigator to learn from an experience, lends itself naturally to the problem of multimodal surface minimization. It has enabled this author to extend the proposed algorithm "SCAN" [7], [8] into a global search strategy.

THE BASIS OF THE ALGORITHM

The algorithm, called "SCAN", proceeds by setting up new starting points in the parameter space, wherefrom local minima are obtained by a unimodal descent technique. For this the Fletcher-Powell gradient method [9] is used. A data structure corresponding to the search process is conveniently presented in a form of a tree, shown in Fig. 1. A tree is denoted as heuristic, since it is generated by heuristic means. Each node XE of a tree, which represents one starting point in the parameter space, has a pair node, XM, which corresponds to a local minimum. Tree nodes are marked by the algorithm as "live" or "dead", depending on whether they are heuristically considered as capable of generating new nodes, or not. A node is "dead" either by constraints on parameters or by certain characteristics of a given surface shape.

Tree nodes are generated using two heuristic rules, whose combined effect may be given in one expression:

$$\underline{XE}(IO)^{(j+1)} = \underline{XE}(N)^{(j)} \pm I \cdot (\underline{XE}(N)^{(j)} - \underline{XM}(N)^{(j)}) \cdot \rho, \quad I = 1, \dots, K$$

where IO is a current tree node which is generated in the tree level (j+1) and generated by a pair $\underline{XE}(N)$, $\underline{XM}(N)$ in the tree level (j); N is a generator node, I is a current extension number and ρ is the reflection coefficient.

The mentioned heuristic rules are called "the rule of reflection" and "the rule of extension". The first rule reads: "in order to obtain a new starting point, located away from the unimodal region already discovered, reflect the concerning minimum $\underline{XM}(N)$ around its starting point $\underline{XE}(N)$. Control the amount of reflection by using a coefficient ρ, whose value is heuristically selected". The extension rule says: "extend the search along the line defined by a generator $\underline{XE}(N)$, $\underline{XM}(N)$, by creating altogether 2K new starting points, K in the positive and K in the negative direction from $\underline{XE}(N)$, $\underline{XM}(N)$. Choose the value for K heuristically".

The extension rule is similar to the "pattern more" in the direct search method by Hooke and Jeeves [10]. The idea is that having once established a successful direction, it is worth trying it several times before it is abandoned.

The process of a tree growth ends when there are no more live pairs among generated nodes, that is, the routine has found all minima which it is capable of detecting from that particular starting point and a particular choice for K and ρ, which control the tree growth.

DESCRIPTION OF THE PROPOSED GLOBAL STRATEGY

The proposed strategy may be described as the five stage cyclic process which is shown in Fig. 2. The process starts by a stage named START, which involves a heuristic choice of a starting point in a parameter space and the implementation of a descent technique which finds a local minimum. In the second stage, named ASSUMPTIONS, new starting points are generated by applying the extension rule to the first generator pair, XE(1) and XM(1), see Fig. 1. From these starting points new minima are detected which together with the corresponding starting points form new generating pairs. Before further applications of the descent technique are attempted, a newly created group of pairs is evaluated.

The purpose of the evaluation, which is performed in the 3rd stage, see

Fig. 2, is to establish whether all obtained pairs are capable of producing further useful starting points. Those which are capable are marked "live", useless pairs are marked "dead" and are removed from further local minimizations. In a heuristic tree, see Fig. 1, this is reflected by end-points from which no branches are extended. After the EVALUATIONS stage the process proceeds if there are remaining live pairs, otherwise the END stage has been reached.

During one cycle of "SCAN" some or all available minima may be found. It may be desirable to repeat the whole cycle from a different starting point, again heuristically selected, with an objective of a possible detection of those minima which were not found during the first cycle. The knowledge of the program performance during any one cycle provides very useful clues as how to select a new starting point or how to adjust the values of K and ρ for the next run. The availability of this heuristic knowledge is denoted by a broken feedback line connecting the END and the START. The line is broken because at present the closure is effected manually, either in an interactive or a batch processing mode, see Fig. 2, stage 5. An automatic feedback can be mechanized and is an object of a further study.

The global aspect is handled by this strategy in two levels. First, by several trial runs a recognition of the surface is obtained, which determines the promising values for K and ρ. Second, several starting points may be attempted. Because the choice is heuristic it immediately becomes obvious how further improvements can be made. Numerous experiments were undertaken using the test function [11] which is capable of simulating various characteristics of a multimodal surface shape. Usually after only two or three trial runs it was possible to find all specified minima in one single run.

PERFORMANCE OF THE METHOD - A LEARNING BASIS OF THE GLOBAL SEARCH

The method's performance may be described in terms of the "passive" tree, the "oscillator" tree and the "active" tree. The "passive" heuristic tree is a tree which does not develop, which is recognized by the fact that one or none new minimum is found in the level 1, see Fig. 1; owing to the way the procedure evaluates capable generator pairs, a passive tree dies immediately. The "oscillator" tree finds exactly two minima or one minimum and one saddle point along the same search direction and keeps finding them all the time until some stop criterion is effected. The "active" tree is the one which survives long enough to locate most (often all) available minima in one single run. It is capable of this because it catches enough different minima in the level 1, which enables the search directions created to span the space progressively and densely. Both the progress and the density can be controlled by K and ρ constants, respectively. When the optimal values for K and ρ are found for the surface which is optimized, a very efficient result follows, which is seen by the fact that not many tree nodes are needed to achieve a complete success.

The progress and the density of the tree development are well managed by choosing $3 \leq K \leq 8$ and $1.1 \leq \rho \leq 1.6$. The reasoning behind these numbers is as follows: for a small value of ρ ($\rho \leq 0.4$), the algorithm is likely to spend too much time in the same unimodal region, but for a large value of ρ ($\rho \geq 2.0$) some minima may be missed. By taking fewer extension points ($K \leq 2$) a tree is not given much chance to develop but for the first trial runs this may be preferred as such trees die fast, nevertheless giving a useful information on how to modify K, ρ or the starting point. If the minima cluster is dense, a moderate K ($K = 5$) is preferred to a large K ($K = 10$), but it is the other way around for very scarcely scattered minima.

In any case, suitable values for these numbers can be found after brief experimenting with the surface which is optimized. With the examples that were tried the optimal values were often found close to $K=6$ and $\rho = 1.3$. A good guideline might be to start off with these values and concentrate on changing the starting point, until a dramatically good result is obtained. Further improvements may be achieved by varying K and ρ slightly, at a time.

SOME RESULTS AND THE SUMMARY

Numerous tests were undertaken using the scalar function [11] which simulates a multimodal surface. The function $f(\underline{x})$ is defined as

$$\frac{1}{f(\underline{x})} = \sum_{i=1}^{m} \frac{1}{f^i(\underline{x})} + \frac{1}{f^o(\underline{x})},$$

where

$$f^i(\underline{x}) = (\underline{x} - \underline{a}^i)^2 + c^i,$$

$$f^o(\underline{x}) = c^o,$$

$c^i << c^o$, m is the number of minima.

This type of function allows optional specification of position $\underline{a}^i$ and the depth c^i for each of m minima. The plateau surrounding the minima has a value c^o.

Two typical sets of results are shown in Tables I and II. It may be seen in both tables that starting points were commonly selected on the diagonal connecting the origin with an outmost apex (e. g. in Table I, col. 1, row " start at", the coordinates of the point are (1. 5; 1. 5; 1. 5; 1. 5; 1. 5), shortened as 1. 5 all). A glance at Table I, rows for the start, K and ρ, illustrate an important point from the last section: keeping K and ρ at nominal values, while varying the starting point is one easy road to a complete success. Alternatively, by varying one constant, say ρ, while keeping the starting point the same, again the complete success was obtained, as can be seen in cols. 1 and 4 in Table II.

The efficiency of computation is directly related to the highest tree

Table I 10 minima - 5 dimensions; Hypercube defined as $[1.0 \leq x_i \leq 5.0]$, i=5 ; minima placed randomly; Constraints: non-negativity of parameters

Min. no. \ trial no.	1	2	3	4
	Min. was first found at the tree node			
1	1			23
2	--	139	122	24
3	4	10	10	6
4	122	30	277	252
5	--	4	4	5
6	70	36	36	294
7	--	34	34	43
8	6	1	1	1
9	237		317	168
10	3	6	6	7
result	7/10	8/10	9/10	10/10
start at	1.5 all	3.0 all	2.9 all	2.8 all
K ; ρ	4;1.2	4;1.2	4;1.2	4;1.2
tree died at node	289	N. A.	N. A.	N. A.
pages run out at node	N. A.	353	353	N. A.

Table II 15 minima - 15 dimensions; Hypercube defined as $[0.10 \leq x_i \leq 0.20]$, i = 15; Minima placed at some corners of the hypercube; Constraints: non-negativity of parameters

Min. no. \ trial no.	1	2	3	4
	Min. was first found at the node			
1	10	2	5	10
2		4		9
3	9	30	4	3
4	44	20	6	56
5	79	151	41	19
6	45	5	38	31
7	42	49	30	34
8	25	33	13	8
9	8	12	14	2
10	22	13	9	33
11	20	53	10	141
12		6	11	32
13	2	9	8	142
14		84	72	24
15	1	1	1	1
result	12/15	15/15	14/15	15/15
K; ρ	6;1.3	6;1.3	6;1.2	6;1.5
tree died at node	N. A.	N. A.	N. A.	N. A.
pages run out at node	309	N. A.	394	N. A.

node used to achieve a complete result. Here, the surface character affects the situation significantly. At the first sight it might appear odd that the 10 minima-5 dimensional problem required more nodes than 15 minima-15 dimensional problem, see cols. 4 in both tables. The explanation is that randomly placed minima (Table I) interacted more strongly (in a sense of forming sharp-sided saddles), than more evenly distanced minima (Table II), placed at some corners of the hypercube. Because of this, regardless of a lower dimensionality, the number of the tree nodes used, as well as an average number of iterations taken by the descent technique [9] to converge to a minimum in the example in five dimensions, noticably exceeded those needed for the example in fifteen dimensions.

The probability that certain minima may be missed is small for very active trees. In numerous tests undertaken it was possible to locate all specified minima in one run and only once two complementing runs were needed to complete the set. If there is any doubt, starting from a different point and modifying K and ρ systematically, until no further improvement is made in several consecutive runs, should give sufficient insurance in most applications. It is worth bearing in mind that the situation is no better with other methods.

Dimensionality does not affect this strategy. Highly successful experiments were obtained in 15 dimensions and although further dimensions were not attempted, there is no indication that anything critical will develop before the descent technique, used for unimodal minimization, stops performing. The strategy is not affected by dimensionality because it avoids sectionalization of the working space. The price which is presently paid is that some minima are found many times, but the consideration is given as how to minimize this.

A global minimum is just one of the minima found by this approach. Having found them all in a short time (between secs and two mins. on a UNIVAC 1108), the global aspect has been resolved.

Acknowledgements

Dr. R. W. Newcomb is acknowledged for constructive reviewing of manuscript, for encouragement and for making available some funds from his Research Grant AFOSR 70-1910.

Dr. J. Shekel is acknowledged for many fruitful discussions and valid suggestions. His active interest led to the proposition of the test function, which proved a challenge for the method.

References

[1] E. M. Vaysbord: "Convergence of a Method of Random Search", Engineering Cybernetics, No. 3, 1968, pp. 44-48.

[2] I. N. Bocharov, A. A. Feldbaum: "An Automatic Optimizer for the Search for the Smallest of Several Minima", Automatika and Telemekhanika vol. 23, No. 3, March 1962, pp. 289-301.

[3] G. J. McMurtry, K. S. Fu: "A Variable Structure Automaton Used as a Multimodal Searching Technique", IEEE Trans. on Automatic Control, Vol. AC-11, No. 3, July 1966, pp. 379-387.

[4] J. D. Hill: "A Search Technique for Multimodal Surfaces", IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-5, No. 1, January 1969, pp. 2-7.

[5] R. A. Jarvis: "Adaptive Global Search in a Time-Variant Environment Using a Probabilistic Automaton with Pattern Recognition Supervision", IEEE Trans. on Systems Science and Cybernetics, Vol. SSC-6, No. 3, July 1970, pp. 209-217.

[6] M.Oda, B. F. Womack: "A Unified Discussion on Heuristics in Artificial Intelligence", 3rd Asilomar Conference, Dec. 1969, pp. 387-400.

[7] J. Opačić: "SCAN" - a Heuristic Strategy for Solving Problems with Multimodal Criterion Functions", 4th Asilomar Conference, Nov. 1970, pp. 345-350.

[8] J. Opačić: "A tree-structure search algorithm for minimization of multimodal functions, including an identification of stationary points detected", 5th Princeton Conference on Information Sciences and Systems, March 1971 p. 353.

[9] R. Fletcher, M. Powell: "A rapidly convergent descent method for minimization", The Computer Journal, Vol. 6, July 1963, pp. 163-168.

[10] R. Hooke, T. Jeeves: "Direct search solution of numerical and statistical problems", A. C. M, Vol. 8, 1959, pp. 212-229.

[11] J. Shekel: "Test functions for multimodal search techniques", 5th Princeton Conference, March 1971.

FIGURE 1:

A Heuristic Tree.

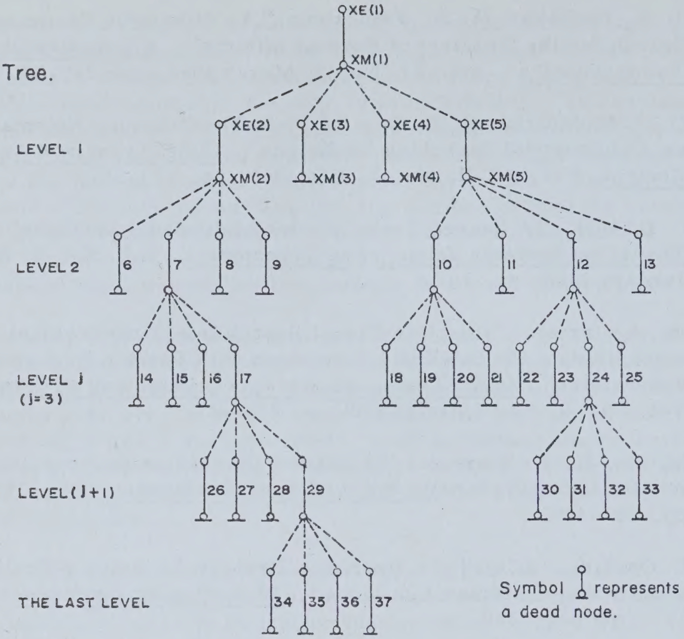
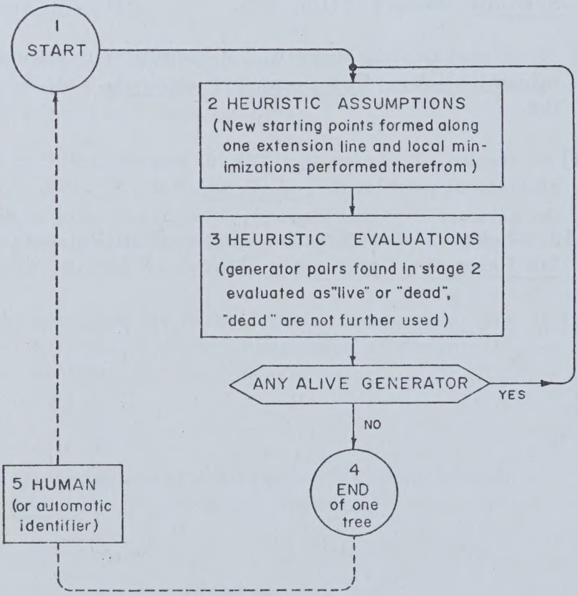

XE(1)
XM(1)
XE(2)
XE(3)
XE(4)
XE(5)
LEVEL 1
XM(2)
XM(3)
XM(4)
XM(5)
LEVEL 2
6
7
8
9
10
11
12
13
LEVEL j
(j=3)
14
15
16
17
18
19
20
21
22
23
24
25
LEVEL (J+1)
26
27
28
29
30
31
32
33
THE LAST LEVEL
34
35
36
37
Symbol represents
a dead node.

FIGURE 2:

A heuristic cycle of the "SCAN" strategy


1
START
2 HEURISTIC ASSUMPTIONS
(New starting points formed along one extension line and local minimizations performed therefrom)
3 HEURISTIC EVALUATIONS
(generator pairs found in stage 2 evaluated as "live" or "dead", "dead" are not further used)
ANY ALIVE GENERATOR
YES
NO
4
END
of one
tree
5 HUMAN
(or automatic
identifier)

EVOLUTIONARY LEARNING OF NEIGHBORHOODS FOR HEURISTIC PROGRAMS AND APPLICATION TO A SEQUENCING PROBLEM*

W.H. KOHLER AND K. STEIGLITZ
Department of Electrical Engineering
Princeton University, Princeton, New Jersey 08540

ABSTRACT

Heuristic programs based on local neighborhood search were proposed by Lin and Reiter and Sherman to generate computer solutions to large-scale traveling salesman problems, and a similar approach has been applied to a variety of other problems. This paper describes a method which incorporates evolutionary learning to direct the selection of neighborhoods for a class of combinatorial problems. The approach consists of two phases:

Phase 1. The evolution of effective neighborhoods for small problems.

Phase 2. The characterization of these neighborhoods in "closed form" for generalization to large problems.

Phase 1 is automated, and Phase 2 presently requires human intervention.

As an example of the overall method, a 5-job one-machine sequencing problem with quadratic tardiness costs is considered. The evolved neighborhood is characterized in terms of simple transformations and the generalized neighborhood structures are then tested on larger problems.

I. INTRODUCTION

Lin [1] and Reiter and Sherman [2] introduced algorithms using local neighborhood search to generate good computer solutions to large-scale traveling salesman problems. This approach has been applied to a variety of other problems as well [3-7]. The fundamental problem in successfully applying local neighborhood search is the selection of an appropriate local neighborhood. This paper describes an evolutionary learning technique for the selection of neighborhoods for

---

* This work was supported by NSF Grant GJ-965, and U.S. Army Research Office - Durham under contract DAHCO4-69-C-0012.

problems whose solution space is isomorphic to a permutation space. When applied to problems where good local neighborhood search algorithms are known, the selected neighborhoods agree with those known to produce good solutions, for example: the 3-change neighborhood for the traveling salesman problem [1] and the pair interchange neighborhood for the problem of assigning economic facilities to locations [3].

II. DEFINITIONS

The local neighborhood search techniques discussed here fall within the following formal framework [7,8].

Definitions:

(1) A combinatorial optimization problem (cop) is a triple (S,X,f) with the following connotation:
   1. The solution space $S = \{s_i\}$, a finite set of feasible solutions.
   2. The parameter space X; each point $x \in X$ represents admissible "data" for the problem.
   3. The cost function $f\colon S \times X \rightarrow R$ where $f(s,x)$ is the cost of solution s with parameter value x.

(2) A globally optimal solution for parameter value x is a feasible solution s^* such that $f(s^*,x) \le f(s,x)$ for all $s \in S$.

(3) A local transformation (lt) is an algorithm which, for each $s \in S$, generates one-by-one a subset of S called the neighborhood of s, denoted N(s).

(4) A pseudorandom start (prs) is an algorithm which produces an element of S with some probability distribution.

(5) A heuristic program (hp) is an algorithm which, using a pseudorandom start and a local transformation N, computes a sequence of trial solutions $s_1, s_2, \ldots, U(s_1)$, such that
   1. s_1, is generated by the pseudorandom starting algorithm;
   2. $s_{i+1} \in N(s_i)$ and $f(s_{i+1},x) < f(s_i,x)$;
   3. $U(s_1)$, the ultimate successor of s_1, is locally optimal in N.

(6) When the hp terminates, the final solution $U(s_1)$ is called lt-opt since it cannot be improved by the application of the lt.

III. APPLICATION TO PERMUTATION PROBLEMS

For the special case when the set of feasible solutions S is isomorphic to the set of permutations of n objects, a neighborhood of a solution, s, can be generated one-by-one by a set of permutation operators $G = \{g_1, g_2, \ldots, g_{|G|}\} \subseteq \{$set of all permutations of n objects$\} \overset{\Delta}{=} S_n$. Selecting a neighborhood is then equivalent to choosing the set of permutation operators G, called the neighborhood generating set. If $s \in S$ then $N(s) = \{g_1(s), g_2(s), \ldots, g_{|G|}(s)\}$, where $g_r(s)$ is the solution obtained by applying permutation g_r to solution s.

In this study the pseudorandom starting algorithm selects a starting solution s by uniformly sampling all feasible solutions. The search for an $s_{i+1} \in N(s_i)$ such that $f(s_{i+1},x) < f(s_i,x)$ is implemented by searching $N(s_i)$ in random order and accepting the first improvement found. The hp algorithm for this scheme is flowcharted as a D-chart [8] in Figure 1. In this go-to-less notation, two control statements are used: an if [condition A] then [statements 1], else [statements 2] indicated by a simple branch; and a while [condition B] do [statements] indicated by a branch to the right of an iteration node "⊗", with condition B written above the branch. The first time condition B fails at the iteration node, execution proceeds downward.

IV. PHASE 1: EVOLUTIONARY SELECTION OF NEIGHBORHOODS

The selection technique attempts to follow the evolutionary principle of selection of the fittest [10]. Random data for small test problems, $n \leq 7$, is generated and permutations are tested to find those which most frequently generate reduced cost solutions. A starting solution, s_i, is obtained using a prs. A permutation $p \in P_j$ is selected at random in proportion to a weight $w_j(p)$, where P_j is the permutation population at generation j. When applied to the current solution s_i, p generates a new solution $p(s_i)$. If $f(p(s_i),x) < f(s_i,x)$, then p becomes a member of the population P_{j+1}, and the current solution is updated, $s_i \leftarrow p(s_i)$. Another permutation is chosen at random from P_j and the process continues until there have been MAXCF consecutive unsuccessful attempts to generate a solution with reduced cost. MAXCF is a parameter which is

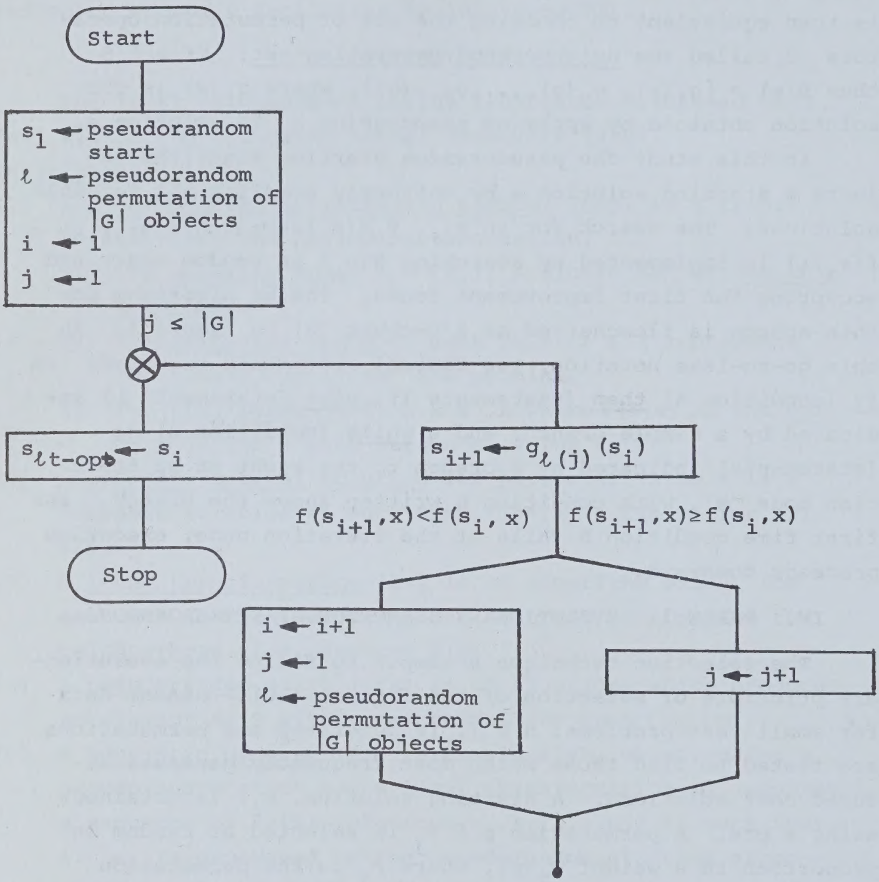

Start
$s_1 \leftarrow$ pseudorandom start
$\ell \leftarrow$ pseudorandom permutation of $|G|$ objects
$i \leftarrow 1$
$j \leftarrow 1$
$j \le |G|$
$s_{\ell t\text{-}opt} \leftarrow s_i$
Stop
$s_{i+1} \leftarrow g_{\ell(j)}(s_i)$
$f(s_{i+1},x)<f(s_i,x)$
$f(s_{i+1},x)\ge f(s_i,x)$
$i \leftarrow i+1$
$j \leftarrow 1$
$\ell \leftarrow$ pseudorandom permutation of $|G|$ objects
$j \leftarrow j+1$

Figure 1: Heuristic Program for Permutation Problems

chosen to allow sufficient sampling of the population P_j. A new starting solution is generated and the process is repeated. Numerous sets of random data are used to eliminate possible data bias. The new population, P_{j+1}, consists of a set of permutations, $P_{j+1} \subseteq P_j$, where each $p \in P_{j+1}$ has a weight, $w_{j+1}(p)$, equal to the number of times it was successful in generating reduced cost solutions. Permutations with zero weight are eliminated. The initial population of permutations, P_o, consists of all n! permutations of n objects, each with equal weight, $w_o(p) = 1$ for all $p \in P_o$. Each new generation of permutations, P_{j+1}, is such that $P_{j+1} \subseteq P_j \subseteq \ldots \subseteq P_1 \subseteq P_0$. If the current generation population has reached the point where it has converged to a small set, it is called the evolved set, P_E, and Phase 1 is complete.

The evolutionary learning algorithm is D-charted in Figure 2. MAXJ, MAXDATA, MAXSTART, and MAXCF are input parameters which control the evolution process and are chosen on the basis of experience with the particular problem.

V. PHASE 2: CHARACTERIZATION AND GENERALIZATION OF THE EVOLVED NEIGHBORHOOD

Experiments have shown that P_E performs very well as a neighborhood generating set, G, for the class of small problems used to get it. Evolution of P_E for large, practical size problems is prohibitive because of the tremendous size, n!, of the initial population P_o. Hence, the problem in Phase 2 is to characterize the evolved set, P_E, in terms of some neighborhood, G, that can be generalized for extension to large-size problems. A generating set, G, is chosen such that $G \cap P_E$ includes those $p \in P_E$ with large weight, $w_E(p)$. G is scored by summing the weight of the permutations in $G \cap P_E$. At the present time, the selection of suitable G's requires human intervention, but it is anticipated that this phase can be automated as well. One possible approach involves generating all permutations of the same conjugacy class [7]. The selection process is completed by choosing G's with high scores. These neighborhoods can then be further analyzed and refined by using them to generate solutions to sets of standard test problems.

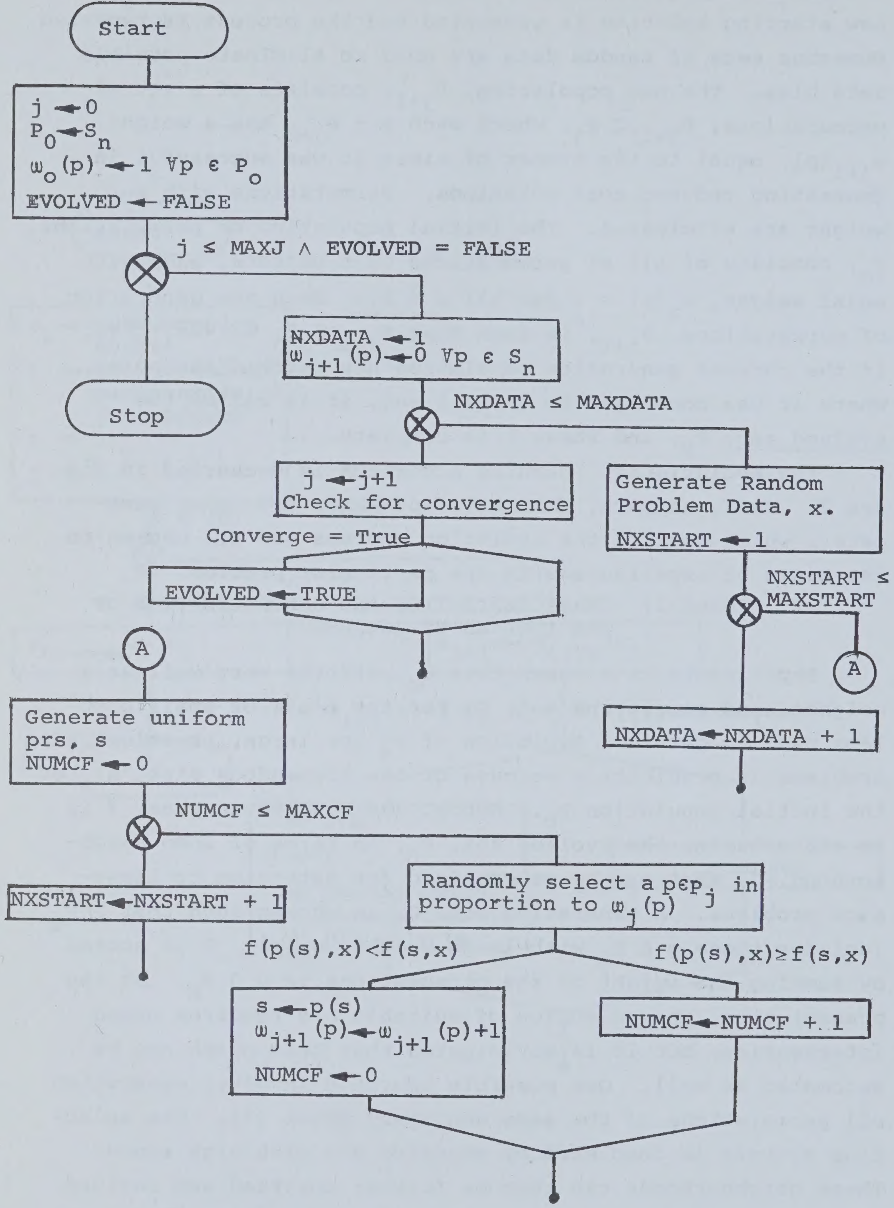

Start
j ← 0
P0 ← Sn
ωo(p) ← 1 ∀p ε Po
EVOLVED ← FALSE
j ≤ MAXJ ∧ EVOLVED = FALSE
Stop
NXDATA ← 1
ωj+1(p) ← 0 ∀p ε Sn
NXDATA ≤ MAXDATA
j ← j+1
Check for convergence
Converge = True
EVOLVED ← TRUE
Generate Random Problem Data, x.
NXSTART ← 1
NXSTART ≤ MAXSTART
A
NXDATA ← NXDATA + 1
A
Generate uniform prs, s
NUMCF ← 0
NUMCF ≤ MAXCF
NXSTART ← NXSTART + 1
Randomly select a pεPj in proportion to ωj(p)
f(p(s),x)<f(s,x)
f(p(s),x)≥f(s,x)
s ← p(s)
ωj+1(p) ← ωj+1(p)+1
NUMCF ← 0
NUMCF ← NUMCF + 1

Figure 2: Evolutionary Learning Algorithm

VI. APPLICATION TO A ONE-MACHINE SEQUENCING PROBLEM WITH TARDINESS COSTS

Using the notation of Conway, Maxwell, and Miller [11], the problem is to sequence a set of n jobs, $J = \{J_1, J_2, \ldots, J_n\}$, with known processing-times, p_i, and due-times, d_i, through a single production facility in such a way as to minimize certain functions of job tardiness. The tardiness, T_i, of job J_i is defined as $T_i = \max(0, C_i - d_i)$, where C_i is the completion time of J_i under the current schedule. Exact techniques for solving special cases of the linear cost case, $\Sigma\, \alpha_i\, T_i$, have been given by McNaughton [12], and a general semi-enumerative algorithm for the linear case was given by Schild and Fredman [13]. Emmons [14] states another semi-enumerative algorithm for linear costs and couples it with a branch-and-bound technique for the case of a convex cost function. Applications of linear programming and dynamic programming techniques were reported by Held and Karp [15] and Lawler [16]. For large-scale problems these programming methods frequently require excessive time and storage. Schild and Fredman [17] developed a procedure for general nonlinear loss functions where the total number of calculations needed to arrive at the optimal schedule is $n(2^{n-1}-1)$. While this technique is a considerable improvement over the enumeration of n! feasible schedules, for n=20 it is already approximately 10^7 and growing rapidly.

One-machine sequencing with quadratic tardiness costs is used to demonstrate the results of applying evolutionary learning to the selection of a heuristic program neighborhood. Random data of the following form was used in the learning and testing phases for n-job problems: processing-times, p_i, are uniformly distributed integers from 1 to 100; and due-times, d_i, are uniformly distributed integers from 1 to 50n. Also associated with each job, J_i, are two cost coefficients, a_{1i} and a_{2i}, where a_{1i} is the linear coefficient and a_{2i} the quadratic coefficient. These coefficients are uniformly distributed integers from 0 to 10. If T_i is the tardiness associated with job J_i under schedule order h, then the total tardiness cost of schedule h is

$$c(h) \triangleq \sum_{i=1}^{n} a_{1i}\, T_i + a_{2i}\, T_i^2 \quad .$$

A. Phase 1

Table 1 shows an evolved population for the case of 5-job sequencing with quadratic tardiness costs. The tenth generation population, P_{10}, was chosen as the evolved population, P_E. P_E consists of 20 distinct permutations.

Table 1: Sample of Neighborhood Learning for 5-Job Sequencing Problems with Quadratic Tardiness Costs.

Population $P_E = P_{10}$

#	Weight	Permutation
1	65	1 2 4 3 5
2	63	1 3 2 4 5
3	61	1 2 3 5 4
4	55	1 3 4 2 5
5	45	1 2 4 5 3
6	44	2 3 1 4 5
7	37	2 1 3 4 5
8	24	2 3 4 1 5
9	22	2 1 3 5 4
10	22	1 3 2 5 4
11	16	2 1 4 3 5
12	10	1 3 4 5 2
13	10	2 3 4 5 1
14	9	1 3 5 4 2
15	4	2 1 5 3 4
16	4	2 1 4 5 3
17	3	2 4 3 1 5
18	3	3 2 4 1 5
19	2	1 4 3 2 5
20	1	1 4 5 2 3

B. Phase 2

Generating sets, G_i, are now chosen such that $G_i \cap P_E$ has a high score. Inspection of the evolved P_E suggested the following G_i: $G_1 = (\text{API})^2$, $G_2 = \text{BSI}$ and $G_3 = G_1 \cup G_2$. $(\text{API})^2$ denotes the set of all adjacent pair interchange permutations of n objects along with all distinct products when taken two at a time. BSI denotes the set of backward single insertion permutations of n objects and consists of the following set of transformations: remove one job from its schedule position and reinsert it anywhere backward of (later than) its original position. Table 2 shows the $(\text{API})^2$ and BSI generating sets for n=5. Using $|G_i|$ to denote the cardinality of set G_i, the neighborhood generating sets are of the following sizes: $|(\text{API})^2| = \binom{n}{2} + (n-2)$, $|\text{BSI}| = \binom{n}{2}$ and $|\text{BSI} \cup (\text{API})^2| = (n-1)^2$. The results of scoring these sets are given in Table 3. All of these neighborhoods have high scores and could be expected to perform well in a heuristic program.

Table 2: Sample Generating Permutation Sets for n=5.

$(API)^2$	BSI
2 1 3 4 5	2 1 3 4 5
1 3 2 4 5	2 3 1 4 5
1 2 4 3 5	2 3 4 1 5
1 2 3 5 4	2 3 4 5 1
3 1 2 4 5	1 3 2 4 5
2 1 4 3 5	1 3 4 2 5
2 1 3 5 4	1 3 4 5 2
2 3 1 4 5	1 2 4 3 5
1 4 2 3 5	1 2 4 5 3
1 3 2 5 4	1 2 3 5 4
1 3 4 2 5	
1 2 5 3 4	
1 2 4 5 3	

Table 3: Sample of Neighborhood Scoring for a 5-Job Sequencing Problem.

Generating Set G_i	$\|G_i\|$	$\|G_i \cap P_E\|$	$\frac{\text{weight } (G_i \cap P_E)}{\text{weight } (P_E)}$
$(API)^2$	13	10	.860
BSI	10	10	.828
BSI ∪ $(API)^2$	16	13	.948

Five standard sets of random data for 5,10,15 and 20-job one-machine sequencing problems were used to test these neighborhoods. The processing-times, due-times and cost coefficients were generated in the previously described manner. Fifty pseudorandom statrs were generated for each data set. Table 4 shows the number of times the best known solution for each data set was found using the three neighborhoods. (Global optimality has been verified for the non-starred data set solutions by an improved version of the Schild and Fredman algorithm [16].) Table 5 gives approximations to the number of schedule cost evaluations per start and execution time required per start on an IBM 360 Model 91. Data and solutions for a sample 5,10,15 and 20-job problem are given in the Appendix. All three neighborhoods performed very well on the random test data. The (API)^2 neighborhood has the lowest probability of finding the best solution, but since the interchange of adjacent jobs taken two at a time changes at most four job completion times, the time required to evaluate the cost change can be kept constant instead of increasing linearly with n. Using this fact, the expected execution time can be decreased considerably over the expression given in Table 5, which did not take advantage of this property.

Table 4: Testing Neighborhoods on Standard Data Sets 50 Random Starts/Set.

Problem		#Times Best Solution Found		
n	Set	(API)^2	BSI	BSI $\cup$ (API)^2
5	1	50	50	50
	2	50	50	50
	3	50	50	50
	4	50	50	50
	5	50	50	50
10	1	22	20	34
	2	50	50	50
	3	39	50	50
	4	50	50	50
	5	28	50	50
15	1	21	50	50
	2*	42	20	43
	3	4	13	13
	4	43	50	50
	5	15	50	50
20	1	2	2	12
	2	5	50	50
	3*	45	50	50
	4*	9	50	50
	5*	19	19	18

Table 5: Approximate Number of Cost Evaluations and Execution Time Per Random Start as a Function of Job Size.

Neighborhood	#Cost Evaluations/Start	Time/Start (seconds)
(API)^2	$.83\ n^{2.26}$	$.0000166\ n^{3.56}$
BSI	$.83\ n^{2.26}$	$.0000123\ n^{3.53}$
BSI $\cup$ (API)^2	$1.6\ n^{2.08}$	$.0000151\ n^{3.65}$

VII. DISCUSSION

The evolutionary learning method represented by Phase 1 has been shown to be a valuable aid in the development of good heuristic programs. It is a step towards the development of heuristic programs which automatically learn how to solve particular problems. The most significant roadblock at this point seems to be in automating Phase 2. This is the problem of generalizing sets of permutations to higher dimensional spaces, and is presently under investigation.

APPENDIX

Sample One-Machine Sequencing Problems.

n=5 Set=1

i	1	2	3	4	5
p_i	95	3	46	72	61
d_i	42	15	38	189	5
a_{1i}	0	6	8	0	7
a_{2i}	5	4	4	6	9

Min. Cost: 232363

Min. Cost Order: 2 5 3 1 4

n=10 Set=1

i	1	2	3	4	5	6	7	8	9	10
p_i	27	23	66	42	60	95	38	37	51	85
d_i	2	354	276	398	160	71	35	217	165	185
a_{1i}	6	1	9	0	0	8	0	8	4	6
a_{2i}	2	7	4	5	3	0	0	10	0	1

Min. Cost: 5364

Min. Cost Order: 1 5 10 8 3 9 2 4 6 7

n=15 Set=1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p_i	12	55	78	34	95	44	46	85	88	36	89	51	90	14	48
d_i	118	659	614	523	671	17	59	490	241	521	548	567	424	548	34
a_{1i}	4	8	4	6	9	7	9	8	0	10	4	4	9	1	0
a_{2i}	5	3	7	5	3	3	9	3	1	8	10	2	9	10	7

Min. Cost: 291927

Min. Cost Order: 15 7 6 1 9 13 4 14 8 10 11 3 12 2 5 n=20 Set=1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
p_i	68	89	93	52	19	66	94	49	83	69	51	75	10	85	29
d_i	124	643	699	101	407	584	872	888	832	130	985	859	66	545	257
a_{1i}	7	3	9	4	8	1	3	4	5	6	9	10	10	1	0
a_{2i}	1	8	6	0	0	8	0	6	1	5	2	3	4	5	2

i	16	17	18	19	20
p_i	94	10	20	41	89
d_i	167	305	964	979	516
a_{1i}	1	4	5	7	7
a_{2i}	0	10	10	9	2

Min. Cost: 3216

Min. Cost Order: 13 10 1 4 15 5 17 20 14 6 2 3 9 12
8 19 18 11 7 16

REFERENCES

1. S. Lin, "Computer Solutions of the Traveling Salesman Problem," BSTJ, pp. 2245-2269, December 1965.

2. S. Reiter and G. Sherman, "Discrete Optimizing," Journal of SIAM, Vol. 13, No. 3, pp. 864-889, September 1965.

3. C.E. Nugent, T.E. Vollmann, and J. Ruml, "An Experimental Comparison of Techniques for the Assignment of Facilities to Locations," Operations Research, Vol. 16, No. 1, pp. 150-173, January-February, 1968.

4. K. Steiglitz, P. Weiner, and D.J. Kleitman, "The Design of Minimum-Cost Survivable Networks," IEEE Trnasactions on Circuit Theory, Vol. CT-16, No. 4, pp. 455-460, November 1969.

5. B. Rothfarb, H. Frank, D.M. Rosenbaum, K. Steiglitz, D.J. Kleitman, "Optimal Design of Offshore Natural Gas Pipeline Systems," Operations Research, Vol. 18, No. 6, pp. 992-1020, November-December 1970.

6. B.T. Bennett and D.C. Gazis, "School Bus Routing by Computer," IBM Research Report RC 3223 (#14797), January 25, 1971.

7. M.J. Krone, "Heuristic Programming Applied to Scheduling Problems," Ph.D. Thesis, Dept. of Electrical Engineering, Princeton University, September 1970.

8. M.J. Krone and K. Steiglitz, "Heuristic Programming: An Empirical Science," Proceedings Third Annual Yale Symposium on Systems Theory, December 1969.

9. J. Bruno and K. Steiglitz, "The Expression of Algorithms by Charts," To be published in Communications of the ACM.

10. H.J. Bremermann, "Optimization Through Evolution and Recombination," in Self-Organizing Systems, (M. Yovits, G. Jacobi, and G. Goldstein, eds.), Spartan Books, Washington, D.C., 1962.

11. R.W. Conway, W.L. Maxwell, and L.W. Miller, Theory of Scheduling, Addison-Wesley, Reading, Mass., 1967.

12. R. McNaughton, "Scheduling with Deadlines and Loss Functions," Management Science, Vol. 6, No. 1, pp. 1-12, October 1959.

13. A. Schild and I.J. Fredman, "On Scheduling Tasks with Associated Linear Loss Functions," Management Science, Vol. 7, No. 3, pp. 280-285, April 1961.

14. H. Emmons, "One-Machine Sequencing to Minimize Certain Functions of Job Tardiness," Operations Research, Vol. 17, No. 4, pp. 701-715, July-August 1969.

15. M. Held and R.M. Karp, "A Dynamic Programming Approach to Sequencing Problems," Journal SIAM, Vol. 10, No. 1, pp. 196-210, March 1962.

16. E.L. Lawler, "On Scheduling Problems with Deferral Costs," Management Science, Vol. 11, No. 2, pp. 280-288, November, 1964.

17. A. Schild and I.S. Fredman, "Scheduling Tasks with Deadlines and Non-Linear Loss Functions," Management Science, Vol. 9, No. 1, pp. 73-81, October 1962.

ON RANDOM SEARCH STRATEGIES FOR PARAMETER OPTIMIZATION

MICHAEL J. WOZNY
School of Engineering
Oakland University
Rochester, Michigan

ABSTRACT

Random search methods offer a practical approach for the identification of parameters in certain applications. In particular the hyperconical random search algorithm is discussed.

I. INTRODUCTION

Although there is a great deal of literature available on the parameter optimization problem, there are very few documented cases on the solution of practical problems in high dimensional parameter spaces. Furthermore, it is certainly clear that high dimensional problems are becoming increasingly important in new application areas of system modeling and optimization. This paper addresses itself to the solution of such problems using random search strategies.

Random search techniques fall into the category of direct search methods, as opposed to gradient related methods. No attempt will be made here to compare gradient and non-gradient methods. The diversity of optimization problems is certainly broad enough so that both techniques offer certain advantages, depending on the class of problems.

One comparison has recently appeared in the literature (1) where it is shown that regardless of dimension the Fletcher-Powell method requires a smaller number of function evaluations than the adaptive step size random method of Schumer and Steiglitz (2), using hyperspherical and hyperellipsoidal surfaces. The authors in (1) also state that it seems intuitively surprizing that "...this minimization can be located more effectively by disregarding all information that can be obtained from the function values and gradients at previous iterations, and proceeding at random, than using this information in some deterministic manner."

It should be noted that surfaces associated with practical optimization problems are generally so irregular that accurate gradients are difficult to compute due to truncation and round-off errors. In such cases one may be better off using no slope information than erroneous slope information!

Two specific parameter identification problems are described in this paper. Both problems are taken from physical applications and are characterized by performance indicies whose surfaces are highly irregular. Reported attempts to solve these problems by gradient related methods have failed; while direct search methods have proved successful. It is hoped that the study of these problems will demonstrate the utility of random search strategies in practical applications.

II. RANDOM SEARCH STRATEGIES

Of the various random search strategies proposed in the literature, four strategies applicable to high dimensional problems are reviewed in this section.

Rastrigin (3,4) proposed a hyperspherical search strategy in which a trial point is chosen from the set of points uniformly distributed on a hypersphere of fixed radius and centered at the previously successful point.

If the trial point is successful, i.e., if it decreases the value of the performance index, then it becomes the center of the hypersphere and a new search is initiated. If the trial point is unsuccessful, it is discarded and a new trial point is chosen from the uniform distribution.

Rastrigin also defined the rate of convergence from one step to the next as the expected improvement along the negative gradient direction. The expected value is taken over the projections onto the gradient of all points on the hypersphere. Clearly, this definition is a true measure of convergence toward the optimum only if the surface representing the performance index is a hypersphere. Otherwise the gradient does not point in the direction of the optimum. Schumer and Steiglitz (2) extended the Rastrigin scheme by allowing the step size, i.e., the radius of the hypersphere, to be variable. The step size was determined by maximizing Rastrigin's measure of the rate of convergence at each iteration.

Bekey (5) used a normal distribution over a hypersphere. However, he added a bias term so that if a successful point is found, a one dimensional search is continued in that direction until no further improvement is obtained. Bremernann's (6) scheme is similar to Bekey's biased search method except that the one dimensional search is performed by fitting a fourth order polynomial and using its minimum as the next trial point.

Wozny and Heydt (7) used a slightly different philosophy. Trial points are taken from a uniform distribution in a hypercone. When a successful point is found, the hypercone apex is moved to the new point, with the cone axis oriented along the line connecting the previous and current successful points. The cone half angle and height are adaptive and are determined by past successes. In essence the hypercone orients itself toward the region most likely to contain a successful point.

The following section discusses the hypercone algorithm in detail. The hypercone algorithm is applied to an example in section IV. Included in this section is a comparison with an adaptive step size hyperspherical strategy.

III. ALGORITHM FOR HYPERCONICAL RANDOM SEARCH

A global minimum is sought for a scalar function P(X) where X is an n dimensional vector. The first two iterates X_1 and X_2 are obtained by a global search such that each coordinate x_i satisfies

$$0 \le |x_i| \le \omega_i$$

X_1 is generated first; X_2 is then sought such that

$$P(X_2) < P(X_1)$$

Each succeeding iterate X_j is sought within a hypercone with an apex at X_{j-1} and an axis parallel to $(X_{j-1} - X_{j-2})$. This hypercone is further defined by a height h and a half angle .

The test points are generated by successively perturbing X_{j-1} in n orthogonal directions. For this reason, an n by n orthonormal matrix is first computed such that the first column is colinear with the cone axis. Each successive column C_k is computed by generating a random vector V, subtracting its component in the direction of each of the preceeding column, and normalizing the result. That is,

$$C_K = \frac{V - \sum_{i=1}^{K-1} (C_i \cdot V) C_i}{\left\| V - \sum_{i=1}^{K-1} (C_i \cdot V) C_i \right\|}$$

The first random perburbation, α_1, is made in the direction of the cone axis (first column of the matrix), and is calculated by

$$\alpha_1 = h(1-y)^n$$

where y is a uniformly distributed random number such that

$$0 \le y \le 1$$

The hypercone radius at this point along the axis is given by

$$r_1 = \alpha_1 \tan\theta$$

The magnitude of each successive perturbation is computed as

$$|\alpha_K| = r_{K-1}\, y^{n-K+1}$$

The sign of this quantity is also randomly selected. After each such perturbation, the radius is revised by

$$r_K = \sqrt{(r_{K-1})^2 - (\alpha_K)^2}$$

After the test point for X_j has been generated, the function P is evaluated at this point. If the test point yields a lower value than $P(X_{j-1})$, it is considered a successful iterate; the cone is realigned, a new orthonormal matrix is computed, the cone height and half angle are modified by

$$h' = s_h h\,, \qquad \theta' = s_\theta \theta$$

and search is begun for X_{j+1}.

If the test point is unsuccessful, however, the cone parameters are modified by

$$h' = u_h h\,, \qquad \theta' = u_\theta \theta$$

and another test point is generated by again perturbing X_{j-1} in the same orthogonal directions. If the number of successive unsuccessful test points exceeds a value I_{max_2}, the hyperconical search yields to a global search to find X_j, after which the hyperconical search is resumed. If, however, the number of successive failures with the global search exceeds a value I_{max_1}, the search is terminated and X_{j-1} is assumed to be the minimum value.

IV. EXAMPLES

Two examples for which gradient related techniques proved inadequate are the Vehicle Bending Mode problem (8) and the Monsanto benchmark problem (9). The first example is connected with modeling the structural rigidity of the frame-body combination in an automobile. As less steel is used in the automobile it becomes an interconnection of flexible highly resonant subsystems. The resulting oscillations lead to premature wear and unwanted vibrations.

The distributed model of the body dynamics consists of a modal frequency response to sinusoidal forces located at discrete spatial locations. For the first five natural modes the model contains sixty-five unknown parameters which must be matched to experimental frequency responses obtained using electrodynamic shakers. The problem is complicated by the fact that the unknown damping ratios for several modes are very small, thus creating steep valleys in the sixty-five dimensional parameter space.

Investigations of this problem showed that gradient realted optimization methods resulted in very poor convergence properties, even when the parameters were initially taken very near the optimum values (8). The problem was eventually solved using a pattern search. At present this problem has not been solved using a random search technique. However, Lawrence and Steiglitz (10) have developed a randomized pattern search which they claim is more effective than the corresponding deterministic method. Consequently, random search looks promising for this problem.

The Monsanto benchmark problem represents a chemical reaction in which the reaction rates are unknown. A model for this chemical reaction consists of seven nonlinear differential equations with twenty-two (for the variable temperature case) unknown parameters. These unknown reaction rates are identified by minimizing a quadratic performance index in the difference between measured concentrations and the calculated concentrations resulting from a given estimate of the reaction rates.

Figure 1 compares the rate of convergence of the hyperconical search and a corresponding adaptive step size hyperspherical search for the constant temperature case (eleven dimensions). It is clear that the directional capabilities of the hyperconical search result in superior convergence rate. Figure 2 shows the tracking capability fo the hyperconical search to continually identify the reaction rate constants (a twenty-two dimensional search) as they vary with temperature. Only two representative reaction rates are shown for each case. Both cases were solved on a CDC 6500 digital computer. Details of the problem statement and problem solution are given in (11). Attempts to solve this problem for the twenty-two dimensional case using gradient techniques by Blevins (12) and Okita (13) were unsuccessful.

V. CONCLUSIONS

It was demonstrated that random search methods offer a practical approach for the identification of parameters in certain applications. In particular, the hypercone random search scheme was discussed. The directional properties of this scheme reveal a faster rate of convergence than an adaptive step size hyperspherical search. Further testing also revealed that the hypercone method appears to converge faster than biased random search. No comparisons are currently available for the Bremermann scheme.

Although random search techniques yield good results, some experience is generally required to "tune" the search parameters of the scheme to the geometry of the problem being solved.

REFERENCES

1. White, L.J. and R.G. Day, "An Evaluation of Adaptive Step-Size Random Search," IEEE Trans. Automat. Contr. (Short Papers), vol. AC-16, pp. 475-478, Oct. 1971.

2. Schumer, M.A. and K. Steiglitz, "Adaptive Step Size Random Search," IEEE Trans. Automat. Contr., vol. AC-13, pp. 270-276, June, 1968.

3. Rastrigin, L. "Extremal Control by the Method of Random Scanning," Avtomatika i Telemekhanika, vol. 21, No. 9, pp. 1264-1271, 1960.

4. Rastrigin, L. "The Convergence of the Random Search Method in the Extremal Control of a Many Parameter System," Avtomatika i Telemekhanika vol. 24, no.11, pp. 1467-1473, Nov. 1963.

5. Bekey, G. and N. Karplus. Hybrid Computation, J. Wiley & Sons., New York, 1968.

6. Bremermann, H. "A Method of Unconstrained Global Optimization," Math. Biosciences, vol.9, pp. 1-15, 1970.

7. Wozny, M. J. and G.T. Heydt, "Hyperconical Random Search," ASME Trans. Journal of Dynamic Systems, Measurement, and Control. March, 1972.

8. Hill, J.C., "Second Progress Report on Flexible Vehicle Parameter Identification," submitted to Vehicles Dynamics Group, Chevrolet Motor Division, General Motors Corp., Warren, Michigan, Sept. 8, 1970.

9. Giese, C., "Determination of Best Kinetic Coefficients of a Dynamic Chemical Process by On-Line Digital Simulation," Simulation, vol.20, 1967.

10. Lawrence, J.P.III, and K. Steiglitz, "Randomized Pattern Search," to be published.

11. Wozny, M.J. and G.T. Heydt, "A Directed Random Search," Purdue Laboratory for Applied Industrial Control. Report No. 32, Purdue University, Lafayette, Indiana, Sept. 1970.

12. Blevins, J.R. "Further Results on the Automatic Identification of Chemical Kinetic Parameters," M.S. Project, Purdue University, Aug. 1969.

13. Okita, S., M.S. Project, Purdue University, Aug. 1970.

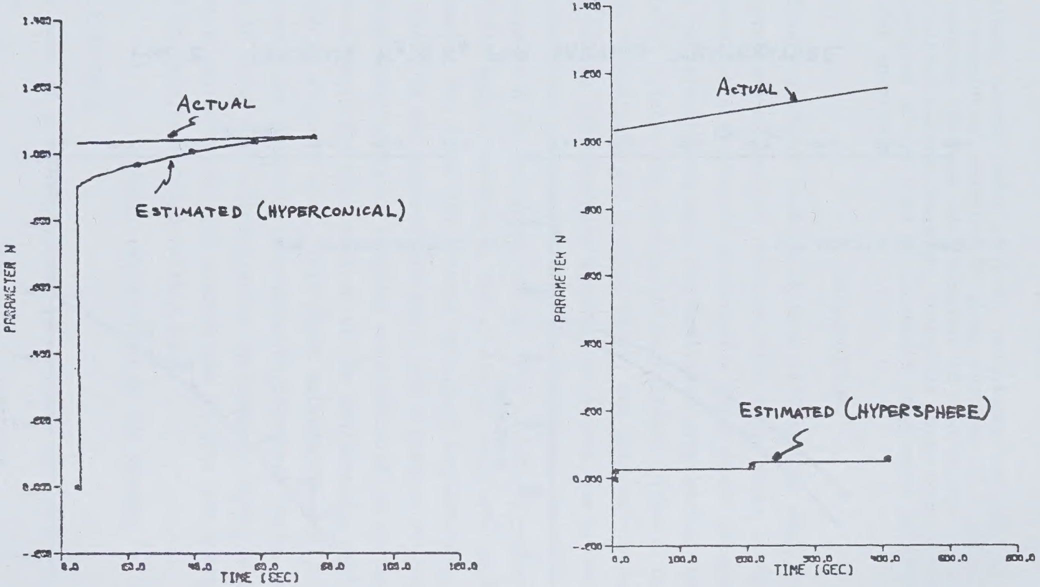


FIG. 1 CONVERGENCE OF K_3

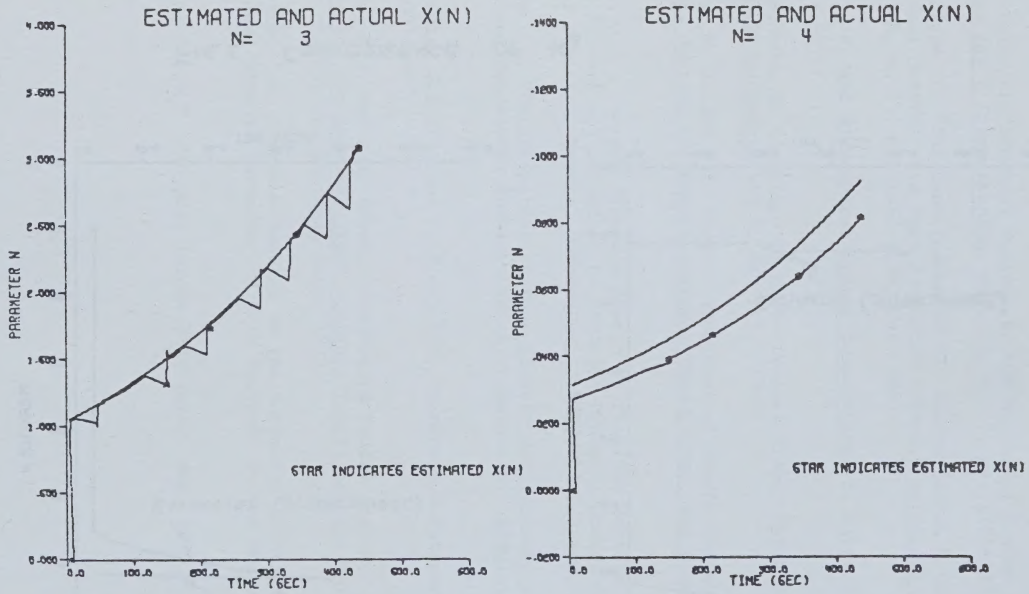


FIG. 2 TRACKING K_3 & K_4 FOR VARYING TEMPERATURE

A SECOND ORDER GENERALIZED REDUCED GRADIENT TECHNIQUE

LEWIS MEIER
Systems Control, Inc.
Palo Alto, California

ABSTRACT

Ordinary second order gradient techniques (as well as all other gradient techniques) work poorly along portions of a solution to a optimum control problem that are near-singular or near-kinky. By application of the Generalized Reduced Gradient Technique these poor convergence problems can be obviated at the cost of increased complexity in solving the state and adjoint equations.

INTRODUCTION

Ordinary gradient techniques have difficulties when applied to singular control problems or control problems with state space constraints. For this reason Mehra and Davis[1] developed a first order generalized reduced gradient technique based upon the nonlinear programming approach of Abadie[2] and Wolfe[3]. Briefly the concept is as follows: Given a function of n+m variables to maximize and n constraints, one selects n variables to be the basis variables and remaining m variables become the optimizing variables. Then one proceeds to optimize with the optimizing variables and satisfy the constraints with the basis variables. For a control problem the constraints become differential equations, and a natural basis and optimizing variables exist-namely, the state and control variables respectively. Normally the natural basis is the appropriate basis for a control problem, but when state-space constraints or singular arcs exist it is preferable to select a different basis for the portions of the trajectories that lie along the constraint or singular arc. This paper extends the work of Mehra and Davis to a second order gradient technique; furthermore, kinky problems, i.e., problems in which discontinuities in derivatives exist, are considered. Kinks are a generalization of constraints in that any constraint can be replaced by penalty function with a kink.

GENERAL DESCRIPTION OF THE METHOD

Problem Statement

For ease of presentation the fixed-final-time problem is considered: maximize

$$J = \int_0^T \ell(x,u,t)dt + \psi[x(T)] \qquad (1)$$

subject to

$$\dot{x} = f(x,u,t), \quad x(0) = x_o \tag{2}$$

where x is the n-dimensional state and u the m-dimensional control. For the present, it is assumed that $\ell(x,u,t)$, $\psi(x,u,t)$, and $f(x,u,t)$ have continuous derivatives of first and second order in x and u.

The Second Order Expansion

Consider the nominal control history $u^\circ(t)$ and the corresponding nominal state history $x^\circ(t)$,

$$\dot{x}^\circ(t) = f^\circ \triangleq f(x^\circ,u^\circ,t), \quad x^\circ(0) = x^\circ, \tag{3}$$

and nominal performance

$$J^\circ = \int_0^T \ell(x^\circ,u^\circ,t)dt + \psi[x^\circ(T)] \tag{4}$$

Note the following identity

$$\int_0^T (\lambda'\Delta\dot{x} + \dot{\lambda}'\Delta x + \Delta x' \, P \, \Delta\dot{x} + \frac{1}{2} \Delta x' \, \dot{P} \, \Delta x)dt = \lambda'\Delta x + \frac{1}{2} \Delta x' \, P \, \Delta x \Big|_0^T \quad , \tag{5}$$

where $\Delta x \triangleq x(t) - x^\circ(t)$, λ and P are an n-dimensional vector and nxn matrix to be defined, and ' indicates matrix transpose. By definition $\Delta x(0) = 0$ and from (2) and (3)

$$\Delta\dot{x} = f(x^\circ+\Delta x,u^\circ+\Delta u,t) - f(x^\circ,u^\circ,t) \tag{6}$$

When (1), (4), (5), and (6) are combined and a Taylor's series expansion to second order made, the result is

$$J \approx J^\circ + \int_0^T [(\dot{\lambda}' + H^\circ_x) \, \Delta x + H^\circ_u \, \Delta u + \frac{1}{2} \Delta x' \, (\dot{P} + H^\circ_{xx} + f^{\circ\prime}_x \, P + Pf^\circ_x)\Delta x + \Delta x'(H^\circ_{xu} + Pf_u)\Delta u + \frac{1}{2} \Delta u' \, H^\circ_{uu} \, \Delta u]dt \tag{7}$$

if

$$H^\circ \triangleq \ell(x^\circ,u^\circ,t) + \lambda' f(x^\circ,u^\circ,t), \tag{8}$$

$$\lambda' \triangleq \psi_x[x^\circ(T),T], \quad P(T) = \psi_{xx}[x^\circ(T),T] \tag{9}$$

Computation of the Generalized Reduced Gradient

Select n of the variables from among the components of x and u and form the basis vector y; the remaining components form the optimization vector z. Equation (7) may be rewritten in terms of $\Delta y = y-y°$ and $\Delta z = z-z°$ as

$$J \approx J° + \int_0^T (a'\Delta y + b'\Delta z + \frac{1}{2}\Delta y' A\Delta y + \Delta y' B\Delta z + \frac{1}{2}\Delta z' C\Delta z)dt, \tag{10}$$

for appropriate a, b, A, B, and C. Now let

$$\Delta z = \delta z + K\Delta y \tag{11}$$

and select λ, P, and K such that

$$a' + b' K = 0 \tag{12}$$

$$A + K'B' + BK + K'C K = 0 \tag{13}$$

$$B' + CK = 0 \tag{14}$$

Then

$$J \approx J° + \int_0^T (b'\delta z + \frac{1}{2}\delta z\ C\ \delta z)dt; \tag{15}$$

b' is known as the generalized reduced gradient (GRG) and C as the generalized reduced second gradient (GRSG).

Solution of the Generalized Adjoint and Riccati Equations

Note that if C is invertible

$$K = -C^{-1} B' \tag{16}$$

satisfies (14) and (12) and (13) become

$$a' - b'C^{-1}B' \tag{17}$$

$$A - BC^{-1}B' = 0 \tag{18}$$

In standard gradient techniques, the state vector x is taken as the basis vector y and the control vector u is used as the optimizing vector z. In this case, $b' = H_u$ and $C = H_{uu}$, (12) becomes the adjoint equation, and (17) becomes the Riccati equation for the linear-quadratic control problem with (6) linearized as state equation and the second derivatives of the Hamiltonian H° as cost matrices. However, if H_{uu} or H_{xx}^{-1} is singular the Riccati equation cannot be solved; furthermore when H_{uu} or H_{xx}^{-1} is near singular, numerical difficulties will occur. In these situations, it is desirable to select some of the state variables as optimizing variables. Equations (17) and (18) will be referred to as the generalized adjoint and Riccati equations.

For simplicity only the case where all the control variables are made basis variables will be considered in detail. In this case with appropriate ordering of x

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad y = \begin{bmatrix} x \\ u \end{bmatrix}, \quad z = x_2$$

(Note that the dimension of x_2 and u are the same.)

Similarly

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}' & P_{22} \end{bmatrix}$$

and K may be partioned into

$$K = [K_{x_1} \ \ K_u] \quad .$$

Equation (16) becomes

$$\begin{aligned} K_{x_1} &= -C^{-1} B_{x_1} , \\ K_u &= -C^{-1} B_u ; \end{aligned} \tag{19}$$

and (17) and (18) become

$$\begin{aligned} -\dot{\lambda}_1' &= H^\circ_{x_1} - (H^\circ_{x_2} + \dot{\lambda}_x') C^{-1} B_{x_1} , \\ 0 &= H^\circ_u - (H^\circ_{x_2} + \dot{\lambda}_2') C^{-1} B_u , \\ A_{x_1x_1} &= B_{x_1}' C^{-1} B_{x_1} \\ A_{x_1u} &= B_{x_1}' C^{-1} B_u , \\ A_{uu} &= B_u' C^{-1} B_u , \end{aligned} \tag{20}$$

where

$$A_{x_1x_1} = \dot{P}_{11} + H_{x_1x_1} + f^{\circ\prime}_{1_{x_1}} P_{11} + P_{11} f^\circ_{1_{x_1}} + f^{\circ\prime}_{2_{x_1}} P_{12}' + P_{12} f^\circ_{2_{x_1}} ,$$

$$A_{x_1u} = H^\circ_{x_1u} + P_{11} f^\circ_{1_u} + P_{12} f^\circ_{2_u} ,$$

$$A_{uu} = H^\circ_{u_1u_1} ,$$

$$B_{x_1}' = \dot{P}_{21} + H^\circ_{x_2x_1} + f^{\circ\prime}_{1_{x_2}} P_{11} + P_{12}' f^\circ_{1_{x_1}} + f^{\circ\prime}_{2_{x_2}} P_{21} + P_{22} f^\circ_{2_{x_1}} ,$$

$$B_u = H^\circ_{x_2 u} + P'_{12} f^\circ_{1_u} + P_{22} f^\circ_{2_u} ,$$

$$C = \dot{P}_{22} + H^\circ_{x_2 x_2} + f^{\circ\prime}_{1_{x_2}} P_{12} + P'_{12} f^\circ_{1_{x_2}} + f^{\circ\prime}_{2_{x_2}} P_{22} + P_{22} f^\circ_{2_{x_2}}$$

The GRG and GRSG are $H^\circ_{x_2} + \dot{\lambda}'_2$ and C.

The set of equation given by (20) has just enough equations to find the unknowns λ_1, λ_2, P_{11}, P_{12}, P_{22}. To numerically solve these equations for the unknowns at times t_k, k=1,... requires use of a numerical differentiation formula that gives $\dot{\lambda}_1(t_k)$ in terms of $\lambda_1(t_j)$, $j \geq k$, etc. Then the equations may be solved recursively starting at $t_k = T$. At each step an implicit equation must be solved; however this would be the case when numerically integrating the ordinary adjoint and Riccati equations using a predictor-corrector integration technique. Note that if B_u is invertible (20) can be converted to the standard adjoint and Riccati equations by solving the final equation for C and substituting into the remaining equations; however the situations in which GRG techniques are useful are those in which B_u is ill-conditioned (i.e., whenever H_{uu} or H^{-1}_{xx} are near singular).

Obtaining a New Nominal Trajectory

The application of the above derivation to optimization is as follows: Given a nominal u°(t), x°(t) a new nominal u*(t), x*(t) with improved performance is found by setting

$$z(t) = z^\circ(t) - c\, C^{-1} b \tag{21}$$

and finding y(t) by solution of (2). The optimum value c* of c is found by a one-dimensional optimization of J; u*(t) and x*(t) are found from the values z*(t) and y*(t) which correspond to c*.

Choice of Basis

The coice of basis for the second order GRG is arbitrary provided: C is invertible, the set of equations (2) is solvable, and (2) can be solved for y(t) given z(t) in a region about y°(t), z°(t). In the first order GRG technique, it is known that the condition that (2) can be solved for y(t) given z(t) in a region about y°(t), z°(t) implies that the generalized adjoint equation is solvable. It is speculated that this condition and the condition that C be invertible imply that (20) is solvable, but this has not been proven. In general the natural basis $y(t) = x(t)$ should be used unless the resulting $C(t)$ (i.e. H_{uu}) is ill-conditioned or $B(t)$ becomes large (as for example if H_{xx}^{-1} is ill-conditioned).

THE LIMITING CASES

Singular Problems

Consider a problem for which $H_{uu}=0$, as, for example, if ℓ and f are linear in u. Order x such that f_{2_u} and $H_{x_2x_2}$ are not singular, then x_2 can be replaced by u in the basis and (20) holds. Since $H_{uu}=0$, $A_{uu}=0$, which implies in turn that B_u, K_u, and A_{x_1u} are zero. Given λ_1 and P_{11}; λ_2, P_{12}, and P_{22} can be found by solving the algebraic equations $H^\circ_u=0$, $B_u=0$, and $A_{x_1u}=0$. In particular λ_2 is always given by

$$\lambda_2' = (\lambda_1' f^\circ_{1_u} + \ell_u)(f^\circ_{2_u})^{-1} \tag{22}$$

and, for example, if H°_{xu} and $f^\circ_{1_{u_1}}$ are zero, P_{12} and P_{22} are zero and $C=H_{x_2x_2}$. In any case the generalized adjoint and Riccati equations reduce to one differential equation for λ_1 and one differential equation for P_{11}.

Kinky Problems

Up to the present time it has been assumed that f and ℓ have continuous derivatives up to second order in x and u; at this point the generalization will be made that discontinuities in derivatives (but not in the function) occur on the p dimensional surface $g(x,u)=0$ where g is a p-vector. Such a surface is known as a kink surface and a very good example of a kink surface results when a constraint of the form $g(x,u)\leq 0$ is replaced by an appropriate kinky penalty function. By proper definition of state and control variables $g(x,u) = g(x_2,u_2)$ where the sum of the dimensions of x_2 and u_2 is p. When the kink surface reduces to $g(u_2)=0$, it is easily handled as follows: Suppose $g[u^\circ_2(t)]=0$; $u_2(t)$ is made part of the optimizing vector z. Let the limits of the derivatives on either side of kink surface be $H^{\circ -}_u$ and $H^{\circ +}_u$. If $H^{\circ -}_u>0$, $H^{\circ +}_u<0$ then $u^*_2(t)$ should equal $u^\circ_2(t)$ since any change in $u_2(t)$ would cause a first order reduction performance. Finally the portion of K corresponding $u_2(t)$ is set to zero.

Of greater difficulty is the general kinky control problem; for a general discussion of such problems see Luenberger.[4] To give some idea of the application of the second order GRG, the case in which the kink is given by $g(x_2)=0$ and u has the same dimension as x_2 will be considered. It is assumed that a $u(t)$, $t_o \le t < t_1$ exists such that if $g[x_2(t_o)]=0$, $g(x_2(t))=0$, $t_o < t \le t_1$; i.e., that the kink can be followed. This problem may be solved as follows: assume $x_2^\circ(t)$ lies on the kink and set $y=x_2(t)$. From the continuity of ℓ and f, $H^\circ_{x_1}$, $H^\circ_{x_1x_1}$, and $H^\circ_{x_1u}$ exist; on the other hand, because of the discontinuities in derivatives of ℓ and f on the kink $H_{x_2x_1}$, $H_{x_2x_2}$ and H_{x_2u} are undefined and therefore so are B_u, C, and P_{22} of (19) and (20). Suppose P_{11} and λ_1 are given, find λ_2 by solution of (22), and let $H^{\circ -}_{x_2}>0$ and $H^{\circ +}_{x_2}<0$, then $x_2^*(t) = x^\circ(t)$ and K=0. The quantity P_{12} is found by solution of the algebraic equation $A_{x_1u}=0$, and λ_1 and P_{11} by solution of the differential equations $-\dot{\lambda}_1' = H^\circ_{x_1}$ and $A_{x_1x_1}=0$.

Note the similarity between this kinky problem and the singular problem. In both cases the basis is the same, λ_2 is found by solution if $H_u=0$, P_{12} is found by solution of $A_{x_1u}=0$, and $K_u=0$. The primary differences are that for the singular problem P_{22} is found by setting $B_u=0$ and K_{x_1} may be non-zero, while for the kinky problem P_{22} is undefined and $K_{x_1}=0$. It is also seen that the singular problem is the limit as $H_{uu}\to 0$ while the kinky problem is the limit as $H^{-1}_{xx}\to 0$.

REFERENCES

1. Mehra, R. K., and Davis, R. E., "A Generalized Gradient Method for Optimal Control Problems with Inequality Constraints and Singular Arcs", Proc. 1971 JACC, pp 144-151. (To appear in IEEE Trans. Auto. Control, Feb. 1972.)

2. Abadie, J., "On the GRG Method for Nonlinear Optimization and Its Applications to Control Problems," EDF Note HI 141/00, August 1969.

3. Wolfe, P., "Methods of Nonlinear Programming: the Reduced Gradient Method", in "Recent Advances in Mathematical Programming", R. L. Grove and P. Wolfe Ed., McGraw-Hill, 1963.

4. Luenberger, D. G., "Control Problems with Kinks", *IEEE Trans. on Auto. Control*, Vol. AC-15, No. 5, October 1970, pp 570-575.

ON CONDITIONS FOR OPTIMALITY IN LEAST pTH APPROXIMATION WITH $p\to\infty$

J.W. BANDLER and C. CHARALAMBOUS
Communications Research Laboratory and
Department of Electrical Engineering
McMaster University
Hamilton, Ontario, Canada

Abstract This paper presents a theoretical discussion of the necessary and sufficient conditions for optimality in generalized nonlinear least pth approximation problems for $p\to\infty$. In the limit the conditions for a minimax approximation are derived, as is to be expected. Numerical examples involving the modelling of a fourth-order system by a second-order model, and quarter-wave transmission-line transformers illustrate the results.

INTRODUCTION

Of great practical importance to network and system designers wishing to approximate a specified response by a network or system response, or desiring to meet or exceed certain design specifications is the optimality of their approximation. A number of workers interested in minimax approximation [1,2,3] have independently arrived at similar conditions for nonlinear minimax approximation problems. These are naturally derivable from the Kuhn-Tucker conditions for a constrained optimum because of the close relationship between nonlinear minimax approximation and nonlinear programming. Bandler [3], in particular, derived the appropriate conditions in a general form suitable for such problems as filter design.

Because of the widespread interest in nonlinear least pth approximation [4], and because of recent results [5] that permit least pth objectives in a more generalized sense to be directly applicable to such problems as meeting or exceeding design specifications as in filter design it is felt that a detailed mathematical discussion of conditions for optimality is highly relevant. Thus, the present paper allows for situations more general than the conventional problem of approximation to a single continuous function on a closed interval.

DEFINITIONS AND ASSUMPTIONS

Define real error functions related to the "upper" and "lower" specifications, respectively, as [5], [6]

$$\begin{aligned} e_u(\underset{\sim}{\phi},\psi) &\triangleq w_u(\psi)(F(\underset{\sim}{\phi},\psi)-S_u(\psi)) \\ e_\ell(\underset{\sim}{\phi},\psi) &\triangleq w_\ell(\psi)(F(\underset{\sim}{\phi},\psi)-S_\ell(\psi)) \end{aligned} \tag{1}$$

where

$F(\underset{\sim}{\phi},\psi)$ is the approximating function (actual response)
$S_u(\psi)$ is an upper specified function (desired response bound)
$S_\ell(\psi)$ is a lower specified function (desired response bound)
$w_u(\psi)$ is an upper positive weighting function
$w_\ell(\psi)$ is a lower positive weighting function
$\underset{\sim}{\phi}$ is a vector containing the k independent parameters
ψ is an independent variable (e.g., frequency or time)

This work was supported by the National Research Council of Canada under grant A7239 and by a Frederick Gardner Cottrell grant from the Research Corporation.

In practice, we will evaluate all the functions at a finite discrete set of values of ψ taken from one or more closed intervals. Therefore, we will let

$$\begin{aligned} e_{ui}(\underset{\sim}{\phi}) &\triangleq e_u(\underset{\sim}{\phi},\psi_i) \qquad && i \in I_u \\ e_{\ell i}(\underset{\sim}{\phi}) &\triangleq e_\ell(\underset{\sim}{\phi},\psi_i) \qquad && i \in I_\ell \end{aligned} \tag{2}$$

where it is assumed that a sufficient number of sample points have been chosen so that the discrete approximation problem adequately approximates the continuous problem. I_u and I_ℓ are appropriate index sets.

Case 1 - Specification Violated In the case when the specification is violated some of the $e_{ui}(\underset{\sim}{\phi})$ or $-e_{\ell i}(\underset{\sim}{\phi})$ are positive. In an effort to meet the specification we can propose the following objective function to be minimized[1]

$$U(\underset{\sim}{\phi}) = \{ \sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^p + \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^p \}^{\frac{1}{p}} \tag{3}$$

where

$$\begin{aligned} J_u &\triangleq \{i \mid e_{ui}(\underset{\sim}{\phi}) \geq 0, \quad i \in I_u\} \\ J_\ell &\triangleq \{i \mid -e_{\ell i}(\underset{\sim}{\phi}) \geq 0, \quad i \in I_\ell\} \end{aligned} \tag{4}$$

and $p>1$.

The larger the value of p the more nearly would we expect the maximum error to be emphasized, since

$$\max_i [e_{ui}(\underset{\sim}{\phi}), -e_{\ell i}(\underset{\sim}{\phi})] = \lim_{p\to\infty} \{ \sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^p + \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^p \}^{\frac{1}{p}} . \tag{5}$$

Case 2 - Specification Satisfied For the case when the specification is satisfied all the $-e_{ui}(\underset{\sim}{\phi})$ and $e_{\ell i}(\underset{\sim}{\phi})$ will be positive. This time, in an effort to exceed the specification by as much as possible, we can propose the following objective function to be minimized[1]

$$U(\underset{\sim}{\phi}) = -\{ \sum_{i\varepsilon I_u} [-e_{ui}(\underset{\sim}{\phi})]^{-p} + \sum_{i\varepsilon I_\ell} [e_{\ell i}(\underset{\sim}{\phi})]^{-p} \}^{-\frac{1}{p}} \tag{6}$$

where we assume

$$\begin{aligned} -e_{ui}(\underset{\sim}{\phi}) &> 0 \qquad && i \in I_u \\ e_{\ell i}(\underset{\sim}{\phi}) &> 0 \qquad && i \in I_\ell \end{aligned} \tag{7}$$

and $p \geq 1$.

The larger the value of p the more nearly would we expect the minimum "error" to be emphasized, since

$$\min_i [-e_{ui}(\underset{\sim}{\phi}), e_{\ell i}(\underset{\sim}{\phi})] = \lim_{p\to\infty} \{ \sum_{i\varepsilon I_u} [-e_{ui}(\underset{\sim}{\phi})]^{-p} + \sum_{i\varepsilon I_\ell} [e_{\ell i}(\underset{\sim}{\phi})]^{-p} \}^{-\frac{1}{p}} . \tag{8}$$

Assumptions It is assumed that a minimum exists in a closed and bounded region of points $\underset{\sim}{\phi}$ and that $e_{ui}(\underset{\sim}{\phi})$ and $e_{\ell i}(\underset{\sim}{\phi})$ are continuous for all i with continuous partial derivatives at least in the neighborhood of the minimum.

---

[1] Since, in this paper, the conditions for optimality are of interest, it is convenient to use slightly different objective functions from those proposed in an earlier work [5].

Then the objective functions proposed are continuous with continuous partial derivatives in the neighborhood of the minimum.

THEOREMS

Theorem 1 At an optimum point $\check{\underset{\sim}{\phi}}_\infty$ for a minimax approximation problem

$$\sum_{i\varepsilon K_u} u_{ui}\, \underset{\sim}{\nabla}\, e_{ui}(\check{\underset{\sim}{\phi}}_\infty) = \sum_{i\varepsilon K_\ell} u_{\ell i}\, \underset{\sim}{\nabla}\, e_{\ell i}(\check{\underset{\sim}{\phi}}_\infty)$$

$$\sum_{i\varepsilon K_u} u_{ui} + \sum_{i\varepsilon K_\ell} u_{\ell i} = 1$$

$$u_{ui} \geq 0 \qquad i \ \varepsilon\ K_u$$

$$u_{\ell i} \geq 0 \qquad i \ \varepsilon\ K_\ell$$

where

$$\underset{\sim}{\nabla} \triangleq \left[\frac{\partial}{\partial\phi_1}\ \frac{\partial}{\partial\phi_2} \cdot \cdot \cdot \frac{\partial}{\partial\phi_k}\right]^T$$

and where $e_{ui}(\check{\underset{\sim}{\phi}}_\infty)$ for $i\ \varepsilon\ K_u$ and $-e_{\ell i}(\check{\underset{\sim}{\phi}}_\infty)$ for $i\ \varepsilon\ K_\ell$ are the equal maxima.

Proof for Case 1 Differentiating equation (3)

$$\underset{\sim}{\nabla} U(\underset{\sim}{\phi}) = \left\{\sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^p + \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^p\right\}^{\frac{1}{p}-1}$$

$$\cdot \left\{\sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^{p-1}\, \underset{\sim}{\nabla}\, e_{ui}(\underset{\sim}{\phi}) - \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^{p-1}\, \underset{\sim}{\nabla}\, e_{\ell i}(\underset{\sim}{\phi})\right\}$$

$$= \left\{\sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^p + \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^p\right\}^{\frac{1}{p}}$$

$$\cdot \left\{\sum_{i\varepsilon J_u} \frac{[e_{ui}(\underset{\sim}{\phi})]^p}{\sum\limits_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^p + \sum\limits_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^p}\ \frac{\underset{\sim}{\nabla}\, e_{ui}(\underset{\sim}{\phi})}{e_{ui}(\underset{\sim}{\phi})}\right.$$

$$\left. - \sum_{i\varepsilon J_\ell} \frac{[-e_{\ell i}(\underset{\sim}{\phi})]^p}{\sum\limits_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi})]^p + \sum\limits_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi})]^p}\ \frac{\underset{\sim}{\nabla}\, e_{\ell i}(\underset{\sim}{\phi})}{[-e_{\ell i}(\underset{\sim}{\phi})]}\right\} \tag{9}$$

The necessary conditions for an optimum of $U(\underset{\sim}{\phi})$ are that

$$\underset{\sim}{\nabla} U(\check{\underset{\sim}{\phi}}_p) = \underset{\sim}{0} \tag{10}$$

where $\check{\underset{\sim}{\phi}}_p$ denotes the optimum parameter vector for particular values of p. Let

$$\mu_i(p) = \frac{[e_{ui}(\check{\underset{\sim}{\phi}}_p)]^p}{\sum\limits_{i\varepsilon J_u} [e_{ui}(\check{\underset{\sim}{\phi}}_p)]^p + \sum\limits_{i\varepsilon J_\ell} [-e_{\ell i}(\check{\underset{\sim}{\phi}}_p)]^p} \tag{11}$$

$$\lambda_i(p) = \frac{[-e_{\ell i}(\check{\underset{\sim}{\phi}}_p)]^p}{\sum\limits_{i\varepsilon J_u} [e_{ui}(\check{\underset{\sim}{\phi}}_p)]^p + \sum\limits_{i\varepsilon J_\ell} [-e_{\ell i}(\check{\underset{\sim}{\phi}}_p)]^p} \tag{12}$$

$$u_{ui} = \lim_{p\to\infty} \mu_i(p) \tag{13}$$

$$u_{\ell i} = \lim_{p\to\infty} \lambda_i(p) \tag{14}$$

Then it is clear that as $p\to\infty$, the necessary conditions for optimality applied to equation (9) yield

$$u_{ui} \begin{cases} = 0 & i \notin K_u \\ \geq 0 & i \in K_u \end{cases} \tag{15}$$

$$u_{\ell i} \begin{cases} = 0 & i \notin K_\ell \\ \geq 0 & i \in K_\ell \end{cases} \tag{16}$$

and

$$\sum_{i\in K_u} u_{ui} + \sum_{i\in K_\ell} u_{\ell i} = 1 \tag{17}$$

Therefore, as $p\to\infty$,

$$\sum_{i\in K_u} u_{ui} \underset{\sim}{\nabla} e_{ui}(\check{\phi}_\infty) = \sum_{i\in K_\ell} u_{\ell i} \underset{\sim}{\nabla} e_{\ell i}(\check{\phi}_\infty) \tag{18}$$

and the theorem is proved.

Proof for Case 2 Following a similar procedure to the one used for Case 1, the same conditions can be derived but in this case we let

$$\mu_i(p) = \frac{[-e_{ui}(\check{\phi}_p)]^{-p}}{\sum\limits_{i\in I_u} [-e_{ui}(\check{\phi}_p)]^{-p} + \sum\limits_{i\in I_\ell} [e_{\ell i}(\check{\phi}_p)]^{-p}} \tag{19}$$

$$\lambda_i(p) = \frac{[e_{\ell i}(\check{\phi}_p)]^{-p}}{\sum\limits_{i\in I_u} [-e_{ui}(\check{\phi}_p)]^{-p} + \sum\limits_{i\in I_\ell} [e_{\ell i}(\check{\phi}_p)]^{-p}} \tag{20}$$

Theorem 2 If the relations in Theorem 1 are satisfied at a point $\check{\phi}_\infty$ and the $e_{ui}(\phi)$ are convex for $i \in I_u$ and the $e_{\ell i}(\phi)$ are concave for $i \in I_\ell$ then $\check{\phi}_\infty$ is optimal.

Proof for Case 1 Assuming the $e_{ui}(\phi)$ are convex and the $-e_{\ell i}(\phi)$ are convex

$$\begin{aligned} e_{ui}((1-\lambda)\phi^1 + \lambda\,\phi^2) &\leq (1-\lambda)e_{ui}(\phi^1) + \lambda\, e_{ui}(\phi^2) \\ -e_{\ell i}((1-\lambda)\phi^1 + \lambda\,\phi^2) &\leq -(1-\lambda)e_{\ell i}(\phi^1) - \lambda\, e_{\ell i}(\phi^2) \end{aligned} \tag{21}$$

for $0 \leq \lambda \leq 1$. Therefore, for $p \geq 1$,

$$\{ \sum_{i\in J_u} [e_{ui}((1-\lambda)\phi^1 + \lambda\,\phi^2)]^p + \sum_{i\in J_\ell} [-e_{\ell i}((1-\lambda)\phi^1 + \lambda\,\phi^2)]^p\}^{\frac{1}{p}}$$

$$\leq \{ \sum_{i\in J_u} [(1-\lambda)e_{ui}(\phi^1) + \lambda\, e_{ui}(\phi^2)]^p$$

$$+ \sum_{i\in J_\ell} [-(1-\lambda)e_{\ell i}(\phi^1) - \lambda\, e_{\ell i}(\phi^2)]^p\}^{\frac{1}{p}}$$

$$\leq (1-\lambda)\{\sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi}^1)]^p + \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi}^1)]^p\}^{\frac{1}{p}}$$

$$+ \lambda\{\sum_{i\varepsilon J_u} [e_{ui}(\underset{\sim}{\phi}^2)]^p + \sum_{i\varepsilon J_\ell} [-e_{\ell i}(\underset{\sim}{\phi}^2)]^p\}^{\frac{1}{p}} \qquad (22)$$

from Minkowski's inequality [7]. Hence,

$$U((1-\lambda)\underset{\sim}{\phi}^1 + \lambda\, \underset{\sim}{\phi}^2) \leq (1-\lambda)\, U(\underset{\sim}{\phi}^1) + \lambda\, U(\underset{\sim}{\phi}^2) \qquad (23)$$

Therefore, $U(\underset{\sim}{\phi})$ is convex and Theorem 2 follows.

Proof for Case 2   Using equation (21) we have, for $p \geq 1$,

$$\{\sum_{i\varepsilon I_u} [-e_{ui}((1-\lambda)\underset{\sim}{\phi}^1 + \lambda\, \underset{\sim}{\phi}^2)]^{-p} + \sum_{i\varepsilon I_\ell} [e_{\ell i}((1-\lambda)\underset{\sim}{\phi}^1 + \lambda\, \underset{\sim}{\phi}^2)]^{-p}\}^{-\frac{1}{p}}$$

$$\geq \{\sum_{i\varepsilon I_u} [-(1-\lambda)e_{ui}(\underset{\sim}{\phi}^1) - \lambda\, e_{ui}(\underset{\sim}{\phi}^2)]^{-p}$$

$$+ \sum_{i\varepsilon I_\ell} [(1-\lambda)e_{\ell i}(\underset{\sim}{\phi}^1) + \lambda\, e_{\ell i}(\underset{\sim}{\phi}^2)]^{-p}\}^{-\frac{1}{p}}$$

$$\geq (1-\lambda)\{\sum_{i\varepsilon I_u} [-e_{ui}(\underset{\sim}{\phi}^1)]^{-p} + \sum_{i\varepsilon I_\ell} [e_{\ell i}(\underset{\sim}{\phi}^1)]^{-p}\}^{-\frac{1}{p}}$$

$$+ \lambda\{\sum_{i\varepsilon I_u} [-e_{ui}(\underset{\sim}{\phi}^2)]^{-p} + \sum_{i\varepsilon I_\ell} [e_{\ell i}(\underset{\sim}{\phi}^2)]^{-p}\}^{-\frac{1}{p}} \qquad (24)$$

Space does not permit a proof of this inequality, which is a counterpart to Minkowski's inequality. Hence,

$$-U((1-\lambda)\underset{\sim}{\phi}^1 + \lambda\, \underset{\sim}{\phi}^2) \geq -(1-\lambda)\, U(\underset{\sim}{\phi}^1) - \lambda\, U(\underset{\sim}{\phi}^2) \qquad (25)$$

Therefore, $U(\underset{\sim}{\phi})$ is convex and Theorem 2 follows.

EXAMPLES

Example 1   The first example is to find a second-order model of a fourth-order system when the input to the system is an impulse for different values of p. The transfer function of the fourth-order system is

$$G(s) = \frac{10(s+4)}{(s+1)(s^2+4s+5)(s+5)} \qquad (26)$$

The transfer function of the second-order model considered is

$$H(s) = \frac{c}{(s+\alpha)^2 + \beta^2} \qquad (27)$$

Therefore, in our case we have

$$S(t) = \frac{3}{20}\exp(-t) + \frac{1}{52}\exp(-5t) - \frac{\exp(-2t)}{65}(3\sin 2t + 11\cos 2t) \qquad (28)$$

$$F(\underset{\sim}{\phi},t) = \frac{c}{\beta}\exp(-\alpha t)\sin\beta t \qquad (29)$$

$$\underset{\sim}{\phi} = [\alpha \quad \beta \quad c]^T \tag{30}$$

$$\psi_i = t_i \qquad 0 \le t_i \le 20 \tag{31}$$

$$e(\underset{\sim}{\phi},t) = F(\underset{\sim}{\phi},\psi) - S(\psi) \tag{32}$$

$$U(\underset{\sim}{\phi}) = \{ \sum_{i=1}^{101} |e(\underset{\sim}{\phi},t_i)|^p \}^{\frac{1}{p}} \tag{33}$$

The above objective function, which corresponds to the situation $S_u=S_\ell=S$ and $w_u=w_\ell=1$ in Case 1, was set up using 101 uniformly spaced points over the range 0 and 20 sec. The values of p used for optimization[2] were 2, 6, 10, 20, 30, 40, 50, 70, 80, 100, 200, 1000 and 10,000. See Figure 1.

The values of $\mu_i(p)$ and $\lambda_i(p)$ were calculated for different values of p and for different values of t. Seven values of t were considered, of which four were the points where the approximately equal extrema occurred. Figure 2 demonstrates the validity of equations (15) and (16).

Example 2 Here, we consider the design of 10Ω to 1Ω quarter-wave transmission-line transformers taking, for convenience, 0.3 as an upper specification for the magnitude of the reflection coefficient ρ over a specified 100% frequency band. The basic problem has been previously defined and analyzed in the context of optimality by Bandler [3]. See also [9] and [10]. The optimum 2-section transformer violates the specification. The optimum 3-section transformer, on the other hand, satisfies the specification. The value of p for both cases in the optimization process[2] was 10,000. 101 uniformly spaced sample points were used.

The maximum values of $e_{ui}(\check{\underset{\sim}{\phi}}_p)$, where $F = |\rho|$ and $S_u = 0.3$, and the frequencies at which they occur are shown in Tables I and II. Using equation (11) and Table I non-zero multipliers in the ratio

3.001 : 2.001 : 1.000

and using equation (19) and Table II non-zero multipliers in the ratio

2.999 : 2.807 : 1.759 : 1.000

can be found. These results may be compared with those of Bandler [3].

CONCLUSIONS

It has been shown [5] that, for violated specifications in generalized least pth approximation, we have a close analogy with minimization of a penalty term designed to bring one closer to the boundary of the feasible region, whereas, for satisfied specifications, we have an analogy with minimization of a penalty term designed to steer a feasible solution deeper into the feasible region. Conventional approximation problems fit into the former category, but conditions of optimality are required for both.

In the limit the conditions for a minimax approximation are obtained as is to be expected. It is felt that new insight into minimax algorithms is gained, since appropriate algorithms might attempt to force these conditions in an iterative manner by using extremely large values of p.

---

[2] Techniques for least pth approximation with extremely large values of p are described elsewhere [8].

TABLE I

OPTIMUM 2-SECTION TRANSFORMER WITH $p=10^4$

(SPECIFICATION VIOLATED - CASE 1)

Maximum of $e_{ui}(\underset{\sim}{\check{\phi}}_p)$	0.12857552	0.12857031	0.12856139
Frequencies (GHz) at which maximum occurs	0.5	1.0	1.5

TABLE II

OPTIMUM 3-SECTION TRANSFORMER WITH $p=10^4$

(SPECIFICATION SATISFIED - CASE 2)

Maximum of $-e_{ui}(\underset{\sim}{\check{\phi}}_p)$	0.10270671	0.10270739	0.10271219	0.10271799
Frequencies (GHz) at which maximum occurs	0.5	0.77	1.23	1.5

REFERENCES

[1] V.F. Dem'yanov, "Sufficient conditions for minimax problems", *Zh. Uychisl. Mat. Fiz.*, vol. 10, pp. 1107-1115, 1970.

[2] J. Medanic, "Solution of the convex minimax problem by the Newton-Raphson method", *Proc. 8th Annual Allerton Conf. on Circuit and System Theory*, Urbana, Ill., pp. 13-22, October 1970.

[3] J.W. Bandler, "Conditions for a minimax optimum", *Proc. 8th Annual Allerton Conf. on Circuit and System Theory*, Urbana, Ill., pp. 23-30, October 1970. See also *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 476-479, July 1971.

[4] G.C. Temes and D.Y.F. Zai, "Least pth approximation", *IEEE Trans. Circuit Theory*, vol. CT-16, pp. 235-237, May 1969.

[5] J.W. Bandler and C. Charalambous, "Generalized least pth objectives for networks and systems", *Proc. 14th Midwest Symposium on Circuit Theory*, Denver, Colo., pp. 10.5.1-10.5.10, May 1971.

[6] J.W. Bandler, "Optimization methods for computer-aided design", *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, pp. 533-552, August 1969.

[7] N.I. Achieser, *Theory of Approximation* (Transl. by C.J. Hyman). New York: Frederick Unger, 1956.

[8] J.W. Bandler and C. Charalambous, "Practical least pth approximation with extremely large values of p", *5th Asilomar Conf. on Circuits and Systems*, Pacific Grove, Calif., November 1971.

[9] J.W. Bandler and P.A. Macdonald, "Cascaded noncommensurate transmission-line networks as optimization problems", *IEEE Trans. Circuit Theory*, vol. CT-16, pp. 391-394, August 1969.

[10] J.W. Bandler and P.A. Macdonald, "Optimization of microwave networks by razor search", *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, pp. 552-562, August 1969.

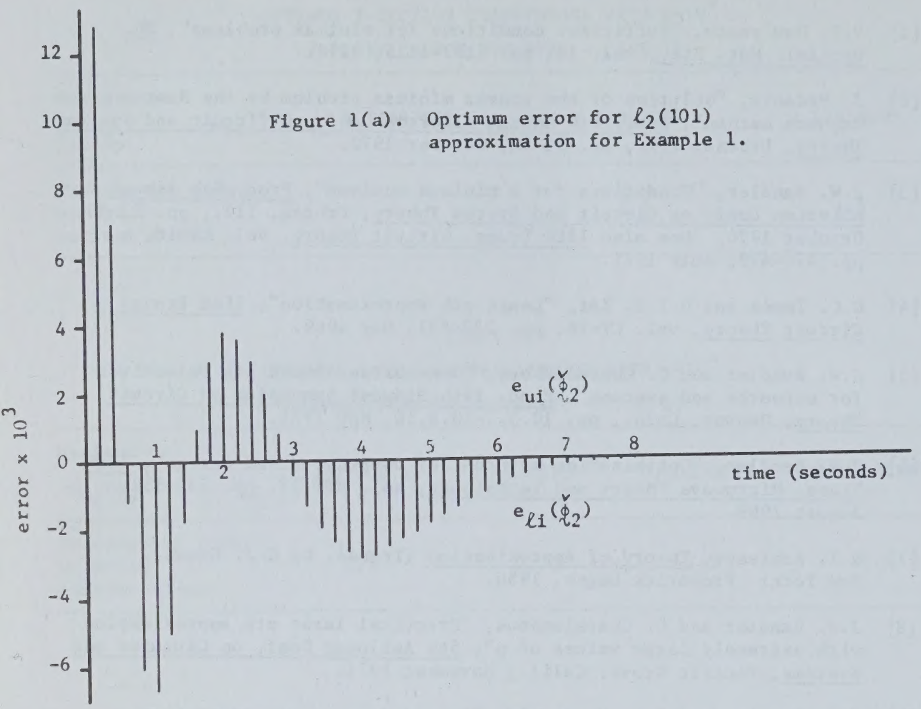


Figure 1(a). Optimum error for $\ell_2(101)$ approximation for Example 1.

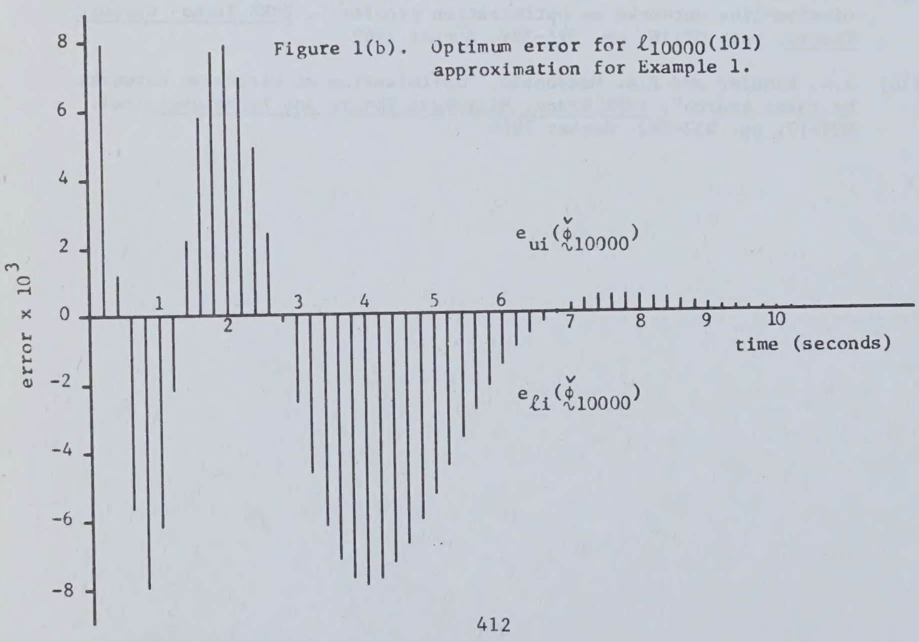


Figure 1(b). Optimum error for $\ell_{10000}(101)$ approximation for Example 1.

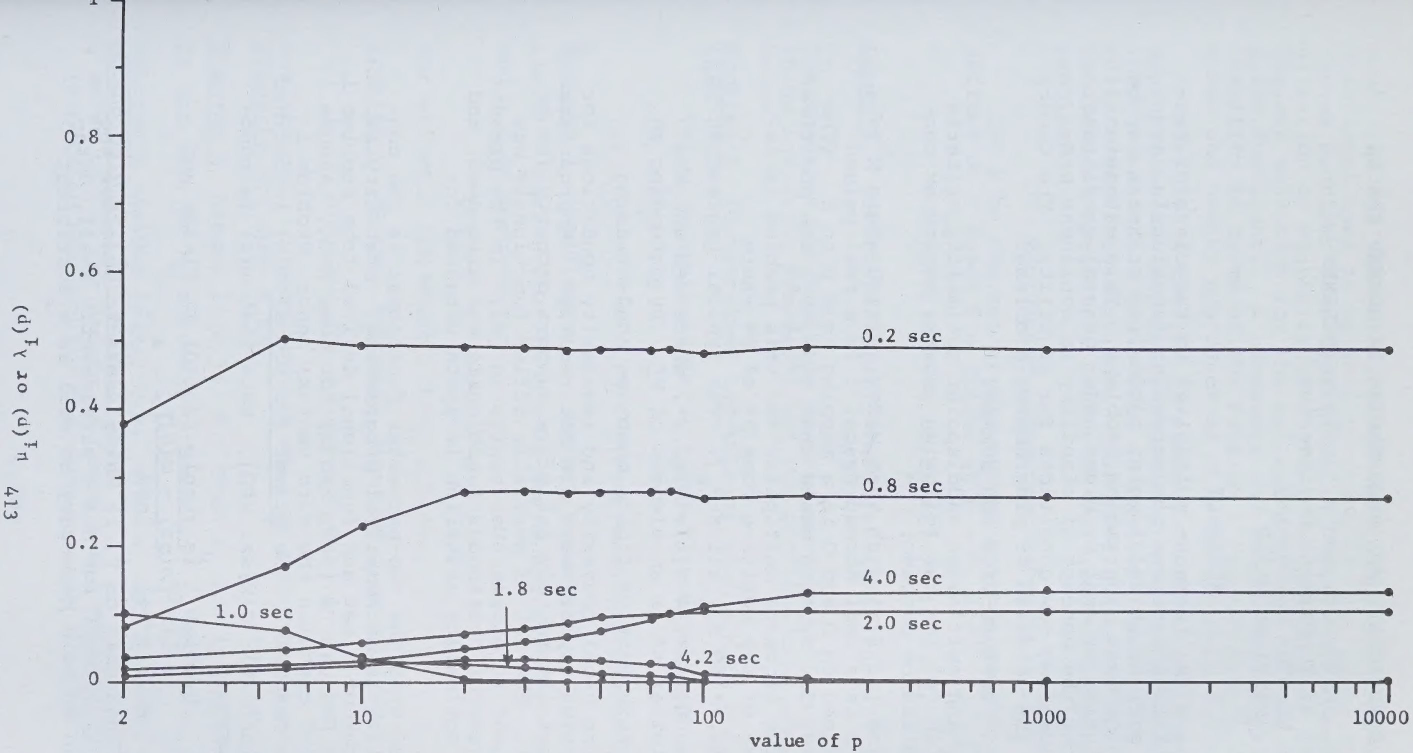


Figure 2. $\mu_i(p)$ or $\lambda_i(p)$, as appropriate, calculated at specified values of time and for certain values of p for Example 1.

TWO-PARAMETER FUNCTIONALS FOR PROGRAMMING IN NORMED SPACES

RICHARD H. SMITH, Jr. AND V. DAVID VANDELINDE
Department of Electrical Engineering
The Johns Hopkins University
Baltimore, Maryland 21218

ABSTRACT

By extending the Lagrange multiplier to the class of two-parameter continuous monotone nondecreasing functionals, new necessary and sufficient saddlepoint optimality criteria may be obtained for nonconvex programming problems. This extension allows the solution of dual problems under general conditions. It also enhances the concept of stability in nonlinear programming and supplies new conditions for stability. The theory is applied to the solution of programming problems.

INTRODUCTION AND SUMMARY

Necessary and sufficient saddlepoint optimality criteria were obtained in [4] for the following general nonconvex constrained optimization problem:

Problem 1 Find inf $f(x)$ such that $x \in \Omega \subset X$, $G(x) \leq 0$ where X is a vector space, Z is a real normed space, f is a real-valued functional defined on Ω and G is a mapping from Ω to Z. (The inequality $z_2 \geq z_1$ for $z_1, z_2 \in Z$ means that $z_2 - z_1 \in P_Z$, the positive cone of Z). The Lagrange multiplier for this problem is definded in terms of the positive cone P_Z^* of Z^* where $P_Z^* \overset{\Delta}{=} \{z^* \in Z^*: \langle z, z^* \rangle \geq 0 \text{ for all } z \in P_Z\}$. The typical Lagrangian is then defined as $L(x,z^*) \overset{\Delta}{=} f(x) + \langle G(x), z^* \rangle$ where $x \in \Omega$ and the Lagrange multiplier z^* is an element of P_Z^*. By extending z^* from P_Z^* to certain sets of finite monotone nondecreasing functionals, the usual convexity and regularity conditions for saddlepoint optimality criteria are not required. Appropriate functionals must alternately satisfy a support property (to be defined later). The set of partially affine functionals was successfully used to obtain the results in [4]. In the present paper, two-parameter continuous functionals are also used, and a saddlepoint optimality criterion is again obtained for Problem 1.

Associated with the two-parameter functional is the concept of stability in mathematical programming. Stability is defined in terms of a set and functional derived from Problem 1. Define the set $\Gamma \subset Z$ as $\Gamma \overset{\Delta}{=} \{z \in Z: z \geq G(x) \text{ for some } x \in \Omega\}$. Assume hereafter that Γ contains the zero vector; hence, Problem 1 will always be feasible. The primal functional $\omega(z)$ is defined on Γ by $\omega(z) \overset{\Delta}{=} \inf\{f(x): G(x) \leq z, x \in \Omega\}$. Note that $\omega(z)$ is monotone nonincreasing.

Definition 1. Problem 1 is stable if $\omega(0)$ is finite and

$$\inf_{\|z\| \neq 0} \lim_{\sigma \to 0+} \left[\frac{\omega(\sigma z) - \omega(0)}{\|\sigma z\|} \right] > -\infty .$$

Since ω is defined only on Γ, it makes sense to consider the above limit only if $\sigma z \in \Gamma$ for all σ sufficiently small. Stability is known to be a necessary as well as a sufficient condition for the existence of the linear Lagrange multiplier functional used in optimality criteria for finite dimensional convex problems (see Geoffrion [1]). In this paper a generalization of stability is shown to be a necessary and sufficient condition for the existence of the two-parameter multiplier functional. A necessary and sufficient condition for stability in terms of the Problem 1 functions $f(x)$ and $G(x)$ is a new and useful byproduct of this development also.

Finally, a theorem is presented which indicates how an appropriate two-parameter functional may be determined. This theorem indicates how a dual to Problem 1 may be solved under quite general assumptions. This result and the corresponding saddlepoint optimality criterion are used to solve programming problems.

EQUIVALENT OPTIMALITY CRITERIA

Let M be the set of all monotone nondecreasing functionals defined on Γ which are finite at $z = 0$, i.e.,

$$M \triangleq \{\ell\colon\ |\ell(0)|<\infty,\ \ell(z_2)\geq\ell(z_1)\ \text{if}\ z_2\geq z_1\}.$$

Lemma 1[+] $\inf\limits_{x\in\Omega}\{f(x)+\ell[G(x)]\}=\inf\limits_{z\in\Gamma}\{\omega(z)+\ell(z)\}$ for all $\ell\in M$.

The theorems which follow in this section will state relationships involving Problem 1 and the infinite dimensional problems below.

Problem 2 (Constrained Lagrangian) Find a $x_0\in\Omega$ and a $\ell_0\in M$ such that

(a) x_0 minimizes $f(x) + \ell_0[G(x)]$ over all $x\in\Omega$

(b) $G(x_0) \leq 0$

(c) $\ell_0[G(x_0)] = \ell_0(0)$. (complementary slackness)

Problem 3 (Saddlepoint) Define the modified Lagrangian as $L(x,\ell) \triangleq f(x) + \ell[G(x)] - \ell(0)$ where $x\in\Omega$ and $\ell\in M$. Find a $x_0\in\Omega$ and a $\ell_0\in M$ such that

$$L(x_0,\ell) \leq L(x_0,\ell_0) \leq L(x,\ell_0)$$

for all $x\in\Omega$, $\ell\in M$ where $|L(x_0,\ell_0)|<\infty$.

Problem 4 (Primal Support) Find an $\ell_0\in M$ such that

$$\omega(0) + \ell_0(0) \leq \omega(z) + \ell_0(z) \tag{1}$$

for all $z\in\Gamma$. Inequality (1) will be referred to as the support property for some $\ell_0\in M$.

Theorem 1 Assume $|f(x_0)|< \infty$. Then x_0 and ℓ_0 solve Problem 2 if and only if x_0 solves Problem 1 and ℓ_0 solves Problem 4.

Theorem 2 Assume $|f(x_0)|< \infty$. Then x_0 and ℓ_0 solve Problem 2 if and only if they solve Problem 3.

[+] Refer to [4] for proofs of Lemma 1 and Theorems 1 and 2. (Proofs do not require the closure of P_z implicit in [4].

BASIC FORM OF TWO-PARAMETER FUNCTIONAL

The two-parameter functional has the form $\ell(z)=k[P(z)]^{\alpha}$ where k and α are scalars with $k \geq 0$ and $0<\alpha\leq 1$. The basic functional P(z) has the properties described below.

Define $d(z_0,K)$ as the distance from the point $z_0 \in Z$ to the convex set $K \subset Z$, i.e.,

$$d(z_0,K) = \inf_{z \in K} \|z_0-z\|.$$

Let N_z denote the closed convex negative cone in z, and let N_z' denote its complement. Finally, define P(z) on Γ with the following properties:

(i) P(z) is monotone nondecreasing and continuous.

(ii) $P(z) = 0$ iff $z \leq 0$; otherwise, $P(z) > 0$.

(iii) $$\sup_{\|z\| \neq 0} \lim_{\sigma \to 0+} \left[\frac{P(\sigma z)}{\|\sigma z\|}\right] < +\infty$$

Since P(z) is defined only on Γ, the limt can be considered only if $\sigma z \in \Gamma$ for all σ sufficiently small.

(iv) $$\inf_{z \in N_z' \cap \Gamma} \left[\frac{P(z)}{d(z,N_z)}\right] > 0 .$$

Functionals satisfying the above properties may be defined with respect to Z. For example, for $Z = E^m$ where $z = (z_1,z_2,\ldots,z_m)$, an appropriate functional is

$P(z) = \sum_{i=1}^{m} [0,z_i]$. For $z = C[a,b]$ where $z = z(t)$, $t \in [a,b]$, P(z) may be defined as $P(z) = \max_{a \leq t \leq b} [0,z(t)]$.

EXISTENCE OF TWO-PARAMETER FUNCTIONAL SATISFYING SUPPORT PROPERTY

The two-parameter functional arises through the following useful generalization of stability.

Definition 2 Problem 1 will be called exponentially stable if $\omega(0)$ is finite and

$$\inf_{\|z\| \neq 0} \lim_{\sigma \to 0+} \left[\frac{\omega(\sigma z)-\omega(0)}{\|\sigma z\|^{\alpha}}\right] > -\infty \text{ for some } \alpha \in (0,1].$$

Note that we must have $\sigma z \in \Gamma$ for all σ sufficiently small.

Exponential stability is a necessary and sufficient condition for the existence of a two-parameter multiplier functional which satisfies the support property.

Theorem 3. Let f(x) be bounded from below on Ω and let $\omega(0)$ be finite. Then there exists a scalar k_0 $(0 \leq k_0 < \infty)$ such that $\ell_0(z) = k_0[P(z)]^{\alpha_0}$ satisfies the support property if and only if Problem 1 is exponentially stable for $\alpha = \alpha_0 \in (0,1]$. In addition, if such a k_0 exists, it solves the dual problem

$$\max_{k \geq 0} \inf_{x \in \Omega} \{f(x) + k[P(G(x))]^{\alpha_0}\}$$ and the maximum equals $\omega(0)$.

(Note that $\ell_0(z)\in M$ by definition of $P(z)$)

Proof: (sufficiency) For $z\in N_z\cap\Gamma$ we have $\omega(0)\le\omega(z)$ because $\omega(z)$ is monotone nonincreasing. Also, $\ell_0(z)=0$ in this case and $\ell_0(0)=0$. Therefore, we immediately have the support property $\omega(0)+\ell_0(0)\le\omega(z)+\ell_0(z)$ for all $z\in N_z\cap\Gamma$.

For $z'\in N_z'\cap\Gamma$, consider $\alpha_0\in(0,1]$, where α_0 is the scalar connected with exponential stability for Problem 1:

$$\frac{\omega(z')-\omega(0)}{[P(z')]^{\alpha_0}} \ge \inf_{z\in N_z\cap\Gamma}\left[\frac{\omega(z)-\omega(0)}{[P(z)]^{\alpha_0}}\right] \tag{2}$$

For each $z\in N_z'\cap\Gamma$, let $\{z_{N_i}(z)\}$ denote a sequence of elements in N_z such that $d(z,N_z)=\lim_{i\to\infty}\|z-z_{N_i}(z)\|$ ($z_{N_\infty}(z)$ may not exist in in general). Since $z_{N_i}(z)\le 0$, we have $z-z_{N_i}(z)\ge z$. Hence, $z-z_{N_i}(z)\in\Gamma$ and $\omega(z-z_{N_i}(z))\le\omega(z)$ by the monotonicity of $\omega(z)$. Consequently, the right-hand side of (2) is greater than or equal to

$$\inf_{z\in N_z'\cap\Gamma}\left[\frac{\omega(z-z_{N_i}(z))-\omega(0)}{[P(z)]^{\alpha_0}}\right]. \tag{3}$$

Let $z_i'(z)=\dfrac{z-z_{N_i}(z)}{d(z,N_z)}$. Note that for $\sigma>0$ arbitrarily small, we may choose i sufficiently large that $1-\sigma\le\|z_i'(z)\|$ and $\dfrac{1-\sigma}{\|z_i'(z)\|}\le 1$ for each $z\in N_z'\cap\Gamma$. By choosing i in this manner for each z, we have that (3) is greater than or equal to

$$\inf_{z\in N_z'\cap\Gamma}\left[(1-\sigma)^{\alpha_0}\cdot\left(\frac{\omega(d(z,N_z)z_i'(z))-\omega(0)}{\|z_i'(z)\|^{\alpha_0}\,d(z,N_z)^{\alpha_0}}\right)\cdot\left(\frac{d(z,N_z)}{P(z)}\right)^{\alpha_0}\right]. \tag{4}$$

Due to the lower boundedness of $f(x)$ on Ω, $\omega(z)$ is bounded from below on Γ; also $\omega(0)<+\infty$. Hence the expression involving ω in (4) is finite if $d(z,N_z)>0$. If $d(z,N_z)\to 0$, this expression remains finite because Problem 1 is exponentially stable. The expression involving $P(z)$ in (4) is finite by property (iv) of $P(z)$. Consequently, (4) is finite. Combining expressions (2) through (4), we have that

$\frac{\omega(z)-\omega(0)}{[P(z)]^{\alpha_0}}$ is greater than or equal to a finite nonpositive scalar $-k_0$ for all $z\in N_z'\cap\Gamma$. Thus, for $\ell_0(z)=k_0[P(z)]^{\alpha_0}$, we have $\omega(0)+\ell_0(0)\leq \omega(z)+\ell_0(z)$ for all $z\in N_z'\cap\Gamma$. Combining this with the result for $z\in N_z\cap\Gamma$, we have the support property for all $z\in\Gamma$.

(Necessity) Assume Problem 1 is not exponentially stable but a finite k_0 exists so that $\ell_0(z)=k_0[P(z)]^{\alpha_0}$ satisfies the support property for some $\alpha_0\in(0,1]$. In this case, we have $\omega(z)-\omega(0)\geq -k_0[P(z)]^{\alpha_0}$ or

$$\frac{\omega(z)-\omega(0)}{[P(z)]^{\alpha_0}}\geq -k_0 \text{ for all } z\in\Gamma, \text{ and } -k_0\leq \inf_{z\in\Gamma}\left[\frac{\omega(z)-\omega(0)}{[P(z)]^{\alpha_0}}\right]$$

$$\leq \inf_{\|z\|\neq 0}\lim_{\sigma\to 0+}\left[\frac{\omega(\sigma z)-\omega(0)}{[P(\sigma z)]^{\alpha_0}}\right]=\inf_{\|z\|\neq 0}\lim_{\sigma\to 0+}\left[\left(\frac{\omega(\sigma z)-\omega(0)}{\|\sigma z\|^{\alpha_0}}\right)\cdot\left(\frac{\|\sigma z\|}{P(\sigma z)}\right)^{\alpha_0}\right].$$

Now $\frac{\|\sigma z\|}{P(\sigma z)}$ is nonzero and positive as $\sigma\to 0+$ by property (iii) of $P(z)$. Hence, the above equals $-\infty$ by the hypothesis of non-exponential stability regardless of the value of α_0. This contradiction proves necessity.

If k_0 exists so that the support property is satisfied, we have $\omega(0)\leq\inf_{z\in\Gamma}\{\omega(z)+k_0[(z)]^{\alpha_0}\}=\inf_{x\in\Omega}\{f(x)+k_0[P(G(x))]^{\alpha_0}\}$ by Lemma 1. On the other hand, $\inf_{x\in\Omega}\{f(x)+k[P(G(x))]^{\alpha_0}\}$ $\leq\inf_{x\in\Omega}\{f(x)+k[P(G(x))]^{\alpha_0}: G(x)\leq 0\}=\inf_{x\in\Omega}\{f(x): G(x)\leq 0\}=\omega(0)$ for any $k\geq 0$. Consequently, k_0 maximizes $\inf_{x\in\Omega}\{f(x)+k[P(G(x))]^{\alpha_0}\}$, and the maximum equals $\omega(0)$. This completes the proof.

Note that the scalar k_0 may be found by a dual maximization without the usual convexity or regularity assumptions which are required in duality theory (see Luenberger [3], p.224).

A CONDITION FOR EXPONENTIAL STABILITY

Stability conditons arise by considering results concerning minimum norm duality (see Nirenberg [2]).

Definition 3 Let K be a convex set in a real normed space Z. The functional $h(z^*)=\sup_{z\in K}\langle z,z^*\rangle$ defined on Z^* is called the support functional of K.

Theorem 4 (Minimum Norm Duality) Let z_0 be a point in a real normed vector space Z and let $d(z_0,K) > 0$ where K is a convex subset of Z having support functional h. Then

$$d(z_0,K) = \max_{\|z^*\|\le 1} [\langle z_0,z^*\rangle - h(z^*)] \tag{5}$$

where the maximum is achieved by some $z_0^* \in Z^*$, $\|z_0^*\|\le 1$.

Lemma 2 The distance function $d(z,N_z)$ is a monotone non-decreasing function of z.

Proof: Since N_z is a cone, the support functional $h(z^*)$ of N_z is finite only for those $z^* \in P_z^*$. Hence, the maximization in (5) is only taken over $z^* \in P_z^*$; and in this case, $h(z^*) \equiv 0$. Therefore, $d(z,N_z) = \max_{\substack{\|z^*\|\le 1\\ z^*\in P_z^*}} \langle z,z^*\rangle = \langle z,z_0^*\rangle$ for some $z_0^* \in P_z^*$, $\|z_0^*\|\le 1$ by Theorem 4. If $z_1 \le z_2$, we have $d(z_1,N_z) = \langle z_1,z_{01}^*\rangle$ (for some $z_{01}^* \in P_z^*$, $\|z_{01}^*\|\le 1$) $\le \langle z_2,z_{01}^*\rangle \le \max_{\substack{\|z^*\|\le 1\\ z^*\in P_z^*}} \langle z_2,z^*\rangle = d(z_2,N_z)$ which proves the lemma.

Theorem 5 Let f(x) be bounded from below on Ω and let $\omega(0)$ be finite. Then Problem 1 is exponentially stable if and only if

$$\inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha_0}}\right] > -\infty \text{ for some } \alpha_0\in(0,1].$$

Proof: (Sufficiency)

$$\inf_{\|z\|\ne 0}\ \lim_{\sigma\to 0+}\left[\frac{\omega(\sigma z)-\omega(0)}{\|\sigma z\|^{\alpha_0}}\right] = \inf_{\|z\|\ne 0}\ \lim_{\sigma\to 0+}\left[\frac{\inf\limits_{x\in\Omega}\{f(x):\ G(x)\le\sigma z\}-\omega(0)}{\|\sigma z\|^{\alpha_0}}\right] \tag{6}$$

(Note that the above limit is considered only if $\sigma z\in\Gamma$ for all σ sufficiently small).

For any x used as an argument in (6), we necessarily have $G(x)\le\sigma z$. Therefore, by Lemma 2 we have $d(G(x),N_z) \le d(\sigma z,N_z) = \inf_{z_N\in N_z}\|\sigma z-z_N\| \le \|\sigma z\|$. In addition, (6) is nonpositive. Consequently, (6) is greater than or equal to

$$\inf_{\|z\|\ne 0}\ \lim_{\sigma\to 0+}\ \inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha_0}} : G(x)\le\sigma z\right]$$

$$\ge \inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha_0}}\right] > -\infty$$ by hypothesis, which proves sufficiency.

(Necessity) Assume

$$\inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha}}\right] = -\infty \tag{7}$$

for all $\alpha\in(0,1]$, but Problem 1 is exponentially stable. Now

$$\inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha}}\right] \geq \inf_{x\in\Omega}\left[\frac{\omega(G(x))-\omega(0)}{d(G(x),N_z)^{\alpha}}\right] \tag{8}$$

by the fact that $\omega(G(x)) \leq f(x)$ for all $x\in\Omega$. For each $x\in\Omega$, let $\{G_{N_i}(x)\}$ represent a sequence of elements in N_z such that $d(G(x),N_z) = \lim_{i\to\infty} \|G(x)-G_{N_i}(x)\|$. Since $G_{N_i}(x) \leq 0$, we have $G(x)-G_{N_i}(x)\geq G(x)$, which implies $G(x)-G_{N_i}(x)\in\Gamma$, and $\omega(G(x)-G_{N_i}(x))\leq \omega(G(x))$ by the monotonicity of $\omega(z)$. Consequently, the right-hand side of (8) is greater than or equal to

$$\inf_{x\in\Omega}\left[\frac{\omega(G(x)-G_{N_i}(x))-\omega(0)}{d(G(x),N_z)^{\alpha}}\right] \tag{9}$$

Define $z_i(x) = \dfrac{G(x)-G_{N_i}(x)}{d(G(x),N_z)}$. Note that for $\sigma > 0$ arbitrarily small, we may choose i sufficiently large that $1-\sigma\leq\|z_i(x)\|$ and $\dfrac{1-\sigma}{\|z_i(x)\|} \leq 1$ for each $G(x)\in\Gamma$. By choosing i in this manner for each $G(x)$, we have that (9) is greater than or equal to

$$\inf_{x\in\Omega}\left[(1-\sigma)^{\alpha}\cdot\left(\frac{\omega(d(G(x),N_z)z_i(x))-\omega(0)}{d(G(x),N_z)^{\alpha}\|z_i(x)\|^{\alpha}}\right)\right] = -\infty \tag{10}$$

by (7). Since $f(x)$ is bounded from below on Ω, $\omega(z)$ is bounded from below on Γ; so (10) can only mean that the infimum occurs at a point where $d(G(x),N_z)\to 0$. Hence, (10) is greater than or equal to

$$\inf_{x\in\Omega}\lim_{\delta\to 0+}\left[(1-\sigma)^{\alpha}\cdot\left(\frac{\omega(\delta z_i(x))-\omega(0)}{\delta^{\alpha}\|z_i(x)\|^{\alpha}}\right)\right]$$

$$\geq \inf_{\|z\|\neq 0}\lim_{\delta\to 0+}\left[(1-\sigma)^{\alpha}\cdot\left(\frac{\omega(\delta z)-\omega(0)}{\|\delta z\|^{\alpha}}\right)\right] = -\infty$$

for all $\alpha\in(0,1]$. This contradiction of exponential stability proves necessity. It should be clear that the infimum in the statement of Theorem 5 is, more specifically, evaluated over the set $S = \{x\in\Omega\colon G(x)\in N_z\}$.

Although the hypothesis of Theorem 5 is difficult to verify in general, it does indicate the usefulness of exponential stability as opposed to simple stability. If f(x) and G(x) are of algebraic form and G(x) contains multiple zeros at solution points, we may expect instability. However, if α is chosen sufficiently small, we may still have exponential stability; and the results in this paper may be used in the solution of problems.

SADDLEPOINT OPTIMALITY CRITERION

Combination of Theorems 3 and 5 gives a necessary and sufficient condition for which $\ell_0(z) = k_0[P(z)]^{\alpha_0}$ with $k_0 \geq 0$, $0 < \alpha_0 \leq 1$ may satisfy the support property.

Corollary 1 Let f(x) be bounded from below on Ω and let $\omega(0)$ be finite. Then $\ell_0(z) = k_0[P(z)]^{\alpha_0}$ satisfies the support property for some scalar $k_0 \geq 0$ if and only if

$$\inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha_0}}\right] > -\infty \text{ for some } \alpha_0 \in (0,1].$$

Corollary 1 is used to state a new necessary and sufficient saddlepoint optimality criterion. For this purpose, define

$$L(x,k,\alpha) \triangleq f(x)+k[P(G(x))]^{\alpha}.$$

Corollary 2 Let f(x) be bounded from below on Ω and let $\omega(0)$ be finite. Assume

$$\inf_{x\in\Omega}\left[\frac{f(x)-\omega(0)}{d(G(x),N_z)^{\alpha_0}}\right] > -\infty \text{ for some } \alpha_0 \in (0,1].$$

Then there exists a finite scalar $k_0 \geq 0$ such that x_0 solves Problem 1 if and only if

$$L(x_0,k,\alpha) \leq L(x_0,k_0,\alpha_0) \leq L(x,k_0,\alpha_0) \text{ and } |L(x_0,k_0,\alpha_0)| < \infty$$

for all $x\in\Omega$, $k \geq 0$, and $\alpha > 0$.

Proof: By definition of P(z), $k[P(z)]^{\alpha} \in M$ for all $k \geq 0$, $\alpha > 0$. Necessity then follows by combining Corollary 1 and Theorem 1 through Problem 4 and Theorems 1 and 2 through Problem 2. For sufficiency, note that the term $k[P(G(x_0))]^{\alpha}$ in $L(x_0,k,\alpha)$ may be made arbitrarily large if $P(G(x_0)) > 0$. Since $L(x_0,k_0,\alpha_0)$ is finite, this circumstance is impossible by the left-hand saddlepoint inequality. This implies that $G(x_0) \leq 0$ by property (ii) of P(z). The fact that $G(x_0) \leq 0$ and the right-hand saddlepoint inequality imply that x_0 and $\ell_0(G(x)) = k_0[P(G(x))]^{\alpha_0}$ solve Problem 2. Therefore, x_0 solves Problem 1 by Theorem 1;

and the corollary is proven.

Applications Find min $f(x) = -x$ such that

$$G(x) = 9x(3x^2-3x+1)-1 \leq 0$$
$$\text{and } x \in [-1,1] = \Omega$$

Ordinarily, a bounded linear functional in P_z^* would be used to provide an alternate method of solution for this problem. However, it may be shown that none exists for which the saddlepoint statement is a necessary optimality criterion. The problem is exponentially stable for $0 < \alpha \leq 1/3$. Hence, a two-parameter functional of the form

$k_0[P(z)]^{1/3}$ will supply an alternate saddlepoint statement by Corollary 2 and Theorem 5. According to Theorem 3, k_0 may be found by solving

$$\max_{k\geq 0} \inf_{x\in\Omega} \{f(x)+k[P(G(x))]^{1/3}\}. \tag{11}$$

We choose $P(G(x))$ as $\max[0,G(x)]$, so (11) becomes

$$\max_{k\geq 0} \inf_{x\in[-1,1]} \left\{\begin{matrix} -x & , G(x) \leq 0 & (x \leq 1/3) \\ -x+k3(x-1/3) & , G(x) \not\leq 0 & (x > 1/3) \end{matrix}\right\}$$

$$= \max_{k\geq 0} \min \left\{\begin{matrix} -1/3 \\ 2k-1, k\leq 1/3 \\ \\ -1/3, (x\sim 1/3),\ k\geq 1/3 \end{matrix}\right\} = \max_{k\geq 0}[2k-1, k\leq 1/3]$$

$$= -1/3 \text{ for } k_0 = 1/3.$$

Note that $x_0 = 1/3$ minimizes the resulting Lagrangian $f(x)+(1/3)\max[0,G(x)]^{1/3}$ and $G(1/3)\leq 0$. Hence, $x_0=1/3$ solves Problem 1 by Theorem 2. Note also that x_0 and $\ell_0(G(x)) = (1/3)\max[0,G(x)]^{1/3}$ may be used in Corollary 2 to imply that x_0 solves Problem 1. Consequently, a solution to Problem 1 is established by various alternate methods. Note finally that the optimal objective value $\omega(0)=f(x_0)= -1/3$ is supplied by the maximum above as stated in Theorem 3.

The techniques used above have also been successfully applied to the solution of infinite dimensional problems. In addition, the two-parameter multiplier functional has lead to the discovery of a two-parameter penalty method for the solution of programming problems (see Smith [5]). The method provides computational advantages for nondifferentiable problems.

REFERENCES

1. Geoffrion, A.M., "Duality in Nonlinear Programming: A Simplified Applications-Oriented Development", SIAM Rev., 13 (1971), pp. 1-37.

2. Nirenberg, L., "Functional Analysis", lectures given in 1960-61, notes by Lesley Sibner, New York Univ., 1961.

3. Luenberger, D.G., Optimization by Vector Space Methods, John Wiley, New York, 1969.

4. Smith, R.H. and VandeLinde, V.D., "Extensions of Lagrange Multipliers in Infinite Dimensional Programming", Proc. Fifth Annual Princeton Conf. Information Sci. Sys., Princeton Univ., Princeton, 1971.

5. Smith, R.H., "Multiplier Functionals for Programming in Normed Spaces", Doctoral dissertation, The Johns Hopkins Univ., Baltimore, 1971.

AN ITERATIVE VECTOR SPACE METHOD FOR EXPONENTIAL APPROXIMATION

R.K. HARMAN AND F.W. FAIRMAN
Department of Electrical Engineering
Queen's University
Kingston, Ontario, Canada.

ABSTRACT

An iterative method is presented for the determination of a weighted sum of exponentials which minimizes the integral squared difference between the weighted sum and the function being approximated. The projection theorem is used to develop a descent method which iteratively adjusts both the weights and exponents of the approximating function. A significant increase in computational efficiency over existing methods is achieved through the use of closed form inverses for the two Gram matrices of interest.

I. INTRODUCTION

Computational difficulties restrict the application of numerous [1]-[5] techniques for best least-squares representation of signals by exponentials. The descent algorithm presented here provides a more efficient approach to a class of these problems. The given quantities for this class of problems are:

$F(s)$, the Laplace transform of a real function $f(t)$,
$F'(s)$, the derivative of $F(s)$ with respect to s.

An approximating function, $f_n(t)$, which minimizes the integral squared error, $E(a_i,b_i)$, between $f(t)$ and $f_n(t)$ is desired.

i.e. $$E(a_i,b_i) = \int_0^\infty |f(t) - f_n(t)|^2 dt \qquad (1)$$

Considering the approximation to be of the form,

$$f_n(t) = \sum_{i=1}^{n} a_i e^{-b_i t} \qquad (2)$$

the a_i and b_i parameters are selected to minimize $E(a_i,b_i)$. In the Laplace domain the approximating function is, $F_n(s)$, where

$$F_n(s) = \sum_{i=1}^{n} \frac{a_i}{(s+b_i)} \qquad (3)$$

The algorithm being developed here is valid provided the following constraints on the elements in the parameter set $\{a_i,b_i\}$; $i = 1,2,\dots,n$ are satisfied.

(a) $\mathrm{Re}\,[b_i] \geq 0$

(b) Whenever $\mathrm{Im}\,[b_i] \neq 0$

then $b_i = b_j^*$ $\quad 1 \leq i \neq j \leq n$

and $a_i = a_j^*$

(c) $b_i \neq b_j \quad 1 \leq i \neq j \leq n$

where $\mathrm{Re}[b_i]$, $\mathrm{Im}[b_i]$, and b_i^* indicate the real part, imaginary part and conjugate of b_i respectively.

A geometric interpretation of the minimization of E using abstract vector spaces was introduced by McDonough and Huggins [1]. This conceptual interpretation is developed in this presentation into a computational technique. For this purpose consider f(t) to be an element in a linear vector space S with subspaces $S_n(b_i)$ and $S_{2n}(b_i)$ generated by $\{e^{-b_i t}\}$ and $\{e^{-b_i t}, te^{-b_i t}\}$. In addition consider the inner product in S to be,

$$< x(t), y(t) > = \int_0^\infty x(t) y^*(t) dt \tag{4}$$

At the start of each iteration in the algorithm the b_i parameters have a particular set of values which determine the orientation of $S_n(b_i)$ and $S_{2n}(b_i)$. The projections of f(t) in these subspaces are used to determine a descent direction for the b_i parameters leading to the next iteration.

II. DEVELOPMENT OF THE METHOD

1) Orthogonal Projection of f(t) onto $S_{2n}(b_i)$

In order to determine descent directions to minimize the error in $S_n(b_i)$ at each stage in the iterative method being developed here, the values of certain parameters $\{\alpha_i, c_i\}$; $i = 1,2,\ldots,n$ are required. These parameters are associated with the orthogonal projection of f(t) onto $S_{2n}(b_i)$. This projection, called $f_{2n}(t)$, is defined as,

$$f_{2n}(t) = \sum_{i=1}^{n} (\alpha_i + c_i t) e^{-b_i t} \tag{5}$$

The α_i and c_i parameters are determined from the normal equations, [6],

$$< f(t) - f_{2n}(t), e^{-b_i t} > = 0$$

$$i = 1,2,\ldots,n$$

$$< f(t) - f_{2n}(t), te^{-b_i t} > = 0$$

which in matrix form is,

$$H \gamma = F_{ab} \tag{6}$$

where; $\gamma = [\alpha_1, c_1, \alpha_2, c_2, \ldots, \alpha_n, c_n]^T$

$F_{ab} = [F(b_1^*), -F'(b_1^*), F(b_2^*), -F'(b_2^*), \ldots\ldots, F(b_n^*), -F'(b_n^*)]^T$

H is a 2n x 2n Hermitian matrix.

The integrals implied by the inner products in H exist and can be evaluated in closed form under the positive real part constraint on the b_i's. More specifically for n = 2, (6) becomes,

$$\begin{bmatrix} \frac{1}{(b_1+b_1^*)}, & \frac{1}{(b_1+b_1^*)^2}, & \frac{1}{(b_2+b_1^*)}, & \frac{1}{(b_2+b_1^*)^2} \\ \frac{1}{(b_1+b_1^*)^2}, & \frac{2}{(b_1+b_1^*)^3}, & \frac{1}{(b_2+b_1^*)^2}, & \frac{2}{(b_2+b_1^*)^3} \\ \frac{1}{(b_1+b_2^*)}, & \frac{1}{(b_1+b_2^*)^2}, & \frac{1}{(b_2+b_2^*)}, & \frac{1}{(b_2+b_2^*)^2} \\ \frac{1}{(b_1+b_2^*)^2}, & \frac{2}{(b_1+b_2^*)^3}, & \frac{1}{(b_2+b_2^*)^2}, & \frac{2}{(b_2+b_2^*)^3} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ c_1 \\ \alpha_2 \\ c_2 \end{bmatrix} = \begin{bmatrix} F(b_1^*) \\ -F'(b_1^*) \\ F(b_2^*) \\ -F'(b_2^*) \end{bmatrix} \qquad (7)$$

2) A Descent Step

General Considerations. A descent step for the a_i and b_i parameters such that $E(a_i,b_i)$ is reduced is now determined by modifying the normal equations (6) into the form,

$$Z\ \delta = -\frac{1}{2}\, g^* \qquad (8)$$

where the a_i and b_i increments are given by,

$$\delta = [\delta_{a_1}, \delta_{b_1}, \delta_{a_2}, \delta_{b_2}, \ldots, \delta_{a_n}, \delta_{b_n}]^T ,$$

and the gradient vector, g by,

$$g = [g_{a_1}, g_{b_1}, g_{a_2}, g_{b_2}, \ldots, g_{a_n}, g_{b_n}]^T$$

having the well known elements, [7]

$$\begin{aligned} g_{a_i} &= \frac{\partial E}{\partial a_i} = -2[F(b_i) - F_n(b_i)] \\ g_{b_i} &= \frac{\partial E}{\partial a_i} = -2a_i[F'(b_i) - F_n'(b_i)] \end{aligned} \qquad (9)$$

and Z is a 2n x 2n positive definite Hermitian matrix to be determined from H and the a_i parameters.

The necessary and sufficient condition for the increments, $\{\delta_{a_i}, \delta_{b_i}\}$ to indicate a descent direction is,

$$\delta^T g < 0 \qquad (10)$$

The proof of this condition in the case of real parameters is well known. [9,p.22]. While one might be tempted to change the transpose in (10) to a conjugate transpose when considering complex parameters this is not the case here. It can be shown, using the conjugacy constraints on the a_i b_i parameters, that descent condition (10) is valid for the approximation being considered. If the δ defined by equation (8) is to satisfy descent condition (10) then,

$$-\frac{1}{2}\, g^{*T}[Z^{-1}]^T g < 0$$

For this reason the increments $\{\delta_{a_i}, \delta_{b_i}\}$ satisfying (8) give rise to a descent direction if and only if Z is positive definite.

Development. The manipulation of the normal equations, (6) into the form of (8) is now considered. Several of the steps in this development will be illustrated using n = 2.

From the definition of $F_n(s)$,(3),

$$\begin{bmatrix} -F_n(b_1^*) \\ F'_n(b_1^*) \\ -F_n(b_2^*) \\ F'_n(b_2^*) \end{bmatrix} = \begin{bmatrix} \frac{1}{(b_1+b_1^*)}, & \frac{1}{(b_2+b_1^*)} \\ \frac{1}{(b_1+b_1^*)^2}, & \frac{1}{(b_2+b_1^*)^2} \\ \frac{1}{(b_1+b_2^*)}, & \frac{1}{(b_2+b_2^*)} \\ \frac{1}{(b_1+b_2^*)^2}, & \frac{1}{(b_2+b_2^*)^2} \end{bmatrix} \begin{bmatrix} -a_1 \\ -a_2 \end{bmatrix} \qquad (11)$$

Since the columns of the matrix in (11) are exactly the first and third columns of the matrix H in (7) these may be added to find,

$$\begin{bmatrix} \frac{1}{(b_1+b_1^*)}, & \frac{1}{(b_1+b_1^*)^2}, & \frac{1}{(b_2+b_1^*)}, & \frac{1}{(b_2+b_1^*)^2} \\ \frac{1}{(b_1+b_1^*)^2}, & \frac{2}{(b_1+b_1^*)^3}, & \frac{1}{(b_2+b_1^*)^2}, & \frac{2}{(b_2+b_1^*)^3} \\ \frac{1}{(b_1+b_2^*)}, & \frac{1}{(b_1+b_2^*)^2}, & \frac{1}{(b_2+b_2^*)}, & \frac{1}{(b_2+b_2^*)^2} \\ \frac{1}{(b_1+b_2^*)^2}, & \frac{2}{(b_1+b_2^*)^3}, & \frac{1}{(b_2+b_2^*)^2}, & \frac{2}{(b_2+b_2^*)^3} \end{bmatrix} \begin{bmatrix} \alpha_1-a_1 \\ c_1 \\ \alpha_2-a_2 \\ c_2 \end{bmatrix} = \begin{bmatrix} F(b_1^*)-F_n(b_1^*) \\ -F'(b_1^*)+F'_n(b_1^*) \\ F(b_2^*)-F_n(b_2^*) \\ -F'(b_2^*)+F'_n(b_2^*) \end{bmatrix} \qquad \ldots(12)$$

More generally (12) can be written as,

$$H\,\zeta = d^* \qquad (13)$$

where

$$\zeta = [\alpha_1 - a_1, c_1, \alpha_2 - a_2, c_2, \ldots, \alpha_n - a_n, c_n]^T$$

$$d = [d_{a_1}, d_{b_1}, d_{a_2}, d_{b_2}, \ldots, d_{a_n}, d_{b_n}]^T$$

$$d_{a_i} = F(b_i) - F_n(b_i)$$

$$d_{b_i} = -F'(b_i) + F'_n(b_i) \qquad (14)$$

From this definition of d and from (9),

$$g = 2\ A\ d \qquad (15)$$

where A is a 2n x 2n diagonal matrix having elements $\{1, -a_1, 1, -a_2, \ldots, 1, -a_n\}$. Substituting (13) into the conjugate of (15),

$$g^* = 2\ A^*\ H\ \zeta \qquad (16)$$

At this point it is important to note that H^{-1} has been developed in closed-form and has been shown to be positive definite. Hence the matrix Z, defined as

$$Z = A^*\ H\ A \qquad (17)$$

is also positive definite. Using (17), (16) can be written as,

$$g^* = 2\ Z\ A^{-1}\ \zeta \qquad (18)$$

To obtain the desired descent equation,

$$Z\ \delta = -\frac{1}{2} g^* \qquad (8)$$

from (18), it is clear that

$$\delta = -A^{-1}\ \zeta \qquad (19)$$

From the definition of ζ, and A, (19) defines,

$$\delta_{a_i} = a_i - \alpha_i \qquad i = 1,2,\ldots,n$$

$$\delta_{b_i} = c_i / a_i$$

This completes the manipulation of (6) into (8).

In order to compute δ one can substitute (17) into (8) and solve, to find

$$\delta = -\frac{1}{2} A^{-1}\ H^{-1}\ A^{*-1} g^* \qquad (20)$$

However, using (15) in (20) a simpler formula results, namely,

$$\delta = A^{-1}\ H^{-1}\ d^* \qquad (21)$$

As a consequence of the foregoing development we are assured that the incremental parameters determined by (21) do indeed move the parameters in a descent direction. Hence, if one defines a new set of parameters by,

$$a_i' = a_i + \beta\ \delta_{a_i} \tag{22}$$

$$b_i' = b_i + \beta\ \delta_{b_i} \tag{23}$$

where the step size, β, is a sufficiently small positive real number, then

$$E(a_i',b_i') < E(a_i,b_i) \tag{24}$$

III. REFINEMENT OF METHOD

1) Orthogonal Projection of f(t) onto $S_n(b_i)$

The unique set of a_i's which minimize $E(a_i b_i)$ for a given set of b_i's is obtained by considering $f_n(t)$ as the orthogonal projection of $f(t)$ onto $S_n(b_i)$.

Applying the projection theorem to determine $f_n(t)$, gives

$$< f(t) - f_n(t), e^{-b_i t} > = 0, \quad i = 1,2,\dots,n \tag{25}$$

Under the positive real part constraints on the b_i's the integrals implied by the above inner products can be evaluated in closed form. This results in the following equation,

$$G\ a = F_b^* \tag{26}$$

where $a = [a_1,a_2,\dots,a_n]^T$

$F_b = [F(b_1),F(b_2),\dots,F(b_n)]^T$

G is an n x n Hermitian matrix having an ijth element,

$$g_{ij} = \frac{1}{(b_i^*+b_j)}$$

The ijth element of G^{-1} is known in closed form, [8] to be

$$\left[G^{-1}\right]_{ij} = \frac{(b_i+b_i^*)(b_j+b_j^*)}{(b_i+b_j^*)} \prod_{\substack{k=1\\k\neq i}}^{n} \left(\frac{b_k^*+b_i}{b_k-b_i}\right) \prod_{\substack{k=1\\k\neq j}}^{n} \left(\frac{b_k+b_j^*}{b_k^*-b_j^*}\right) \tag{27}$$

It is important to note that when the a_i's satisfy (26), the partial derivatives of $E(a_i,b_i)$ with respect to a_i are all zero.

i.e. $$g_{a_i} = 0, \quad i = 1,2,\dots,n \tag{28}$$

From (14), (28) implies,

$$d_{a_i} = 0, \quad i = 1,2,\dots,n \tag{29}$$

This result is used to simplify the computations in the next section.

2) A Two-Stage Descent Step

A computationally efficient two-stage descent step is developed here. This step utilizes the δ_{b_i} incremental parameters determined by (21) in conjunction with the a_i's determined by equation (26).

Given a set $\{b_i\}$, consider the two-stage sequence:

(a) Solve (21) for the δ_{b_i} elements. Determine a new set $\{b_i'\}$ from (23).

(b) Apply (26) to determine the best set $\{a_i''\}$ for the new set $\{b_i'\}$

In comparison to the simultaneous adjustment of the a_i and b_i parameters by the descent step developed earlier , one finds,

$$E(a_i'',b_i') < E(a_i',b_i') < E(a_i,b_i)$$

This further reduction in the error motivates the use of the two-stage descent step, defined by (a) and (b), in the descent algorithm.

The computational aspects of this two-stage descent step can be simplified as follows. Since the a_i's are determined by (26) in stage (b), they need not be determined in stage (a). Hence the δ_{a_i}'s need not be calculated. Furthermore, (29) applies and δ_b can be calculated in stage (a) from the reduced form of equation (21),

$$\delta_b = -A_b^{-1} H_b^{-1} d_b^* \tag{30}$$

where $\delta_b = [\delta_{b_1},\delta_{b_2},\ldots,\delta_{b_n}]^T$

$d_b = [d_{b_1},d_{b_2},\ldots,d_{b_n}]^T$

$A_b = \text{diag}\,\{-a_1,-a_2,\ldots,-a_n\}$

H_b^{-1} is an n x n Hermitian matrix with elements,

$$[H_b^{-1}]_{ij} = [H^{-1}]_{2i,2j}$$

$$[H_b^{-1}] = \frac{(b_i+b_i^*)^2}{(b_i+b_j^*)} \prod_{\substack{k=1\\k\neq i}}^{n} \left(\frac{b_k^*+b_i}{b_i-b_i}\right)^2 \prod_{\substack{k=1\\k\neq j}}^{n} \left(\frac{b_k+b_j^*}{b_k^*-b_j^*}\right)^2 \tag{31}$$

3) Computational Reduction

Due to the similarities in the matrices G^{-1} and H_b^{-1}, which may be observed by examining (27) and (31), on can write

$$[G^{-1}]_{ij} = \frac{v_i v_j^*}{(b_i+b_j^*)} \qquad [H_b^{-1}]_{ij} = \frac{(v_i v_j^*)^2}{(b_i+b_j^*)}$$

where

$$v_i = (b_i+b_i^*) \prod_{\substack{k=1 \\ k\neq i}}^{n} \left(\frac{b_k^*+b_i}{b_k-b_i}\right) \tag{32}$$

Hence the a_i's determined by (26) can be written as,

$$a_i = v_i \sum_{j=1}^{n} \frac{[v_j F(b_j)]^*}{(b_i+b_j^*)} \tag{33}$$

and the δ_{b_i}'s determined by (30) can be written as,

$$\delta_{b_i} = -\frac{v_i^2}{a_i} \sum_{j=1}^{n} \frac{[v_j^2 \, d_{b_j}]}{(b_i+b_j^*)} \tag{34}$$

By using the definition of d_{b_j} given in (14) equation (34) becomes,

$$\delta_{b_i} = \frac{v_i^2}{a_i} \sum_{j=1}^{n} \frac{\{v_j^2 \, [F'(b_j)-F_n'(b_j)]\}}{(b_i+b_j^*)} \tag{35}$$

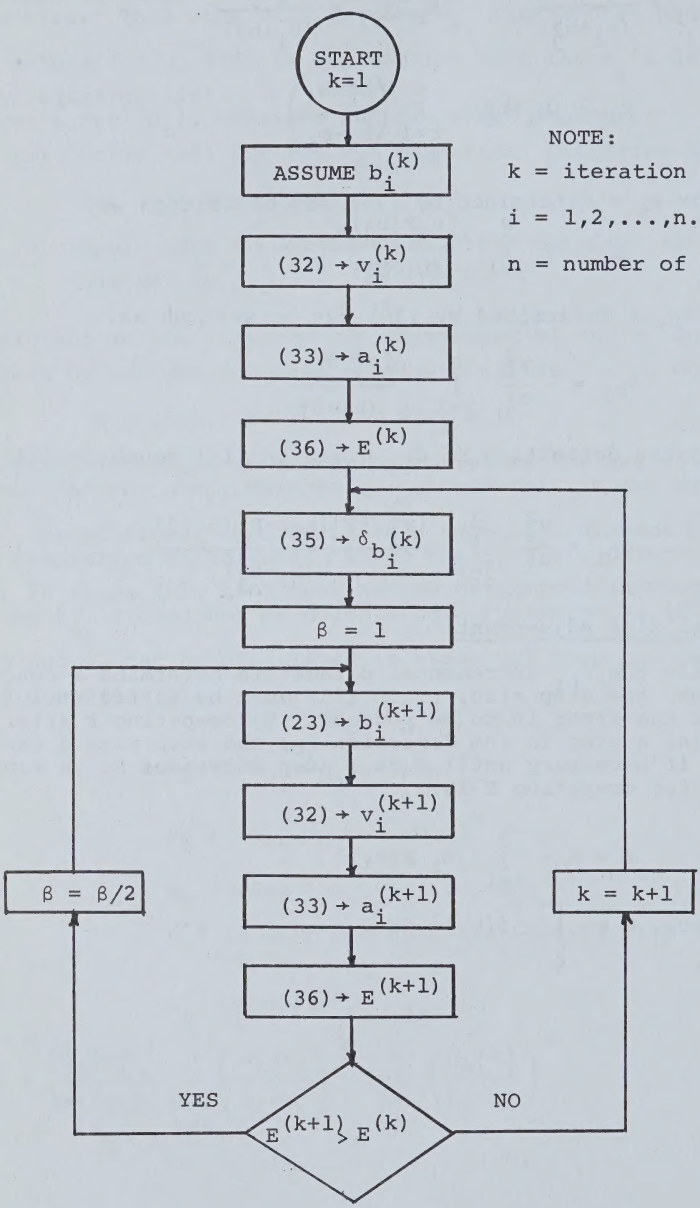
4) Step Size Adjustment

While the δ_{b_i} incremental parameters determine a descent direction, the step size, β, in (23) must be sufficiently small if the error is to be reduced. By computing E after attempting a step in the direction δ_b, the step size β can be reduced if necessary until such a step decreases E. A simple formula for computing E is,

$$E = R - \sum_{i=1}^{n} a_i \, F(b_i) \tag{36}$$

where $R = \int_0^{\infty} [f(t)]^2 \, dt$ (37)

IV THE ALGORITHM


START
k=1
ASSUME $b_i^{(k)}$
(32) → $v_i^{(k)}$
(33) → $a_i^{(k)}$
(36) → $E^{(k)}$
(35) → $\delta b_i^{(k)}$
$\beta = 1$
(23) → $b_i^{(k+1)}$
(32) → $v_i^{(k+1)}$
(33) → $a_i^{(k+1)}$
(36) → $E^{(k+1)}$
$E^{(k+1)} > E^{(k)}$
YES
NO
$\beta = \beta/2$
k = k+1

NOTE:

k = iteration number

i = 1,2,...,n.

n = number of poles

VI. CONCLUDING REMARKS

The computational simplicity of the algorithm presented here facilitates large order approximations. As an example a 20 complex pole approximation to $F(S) = (1-e^{-S})/S$,[1,2], requires only 1.2 seconds per iteration on an IBM 360/50. The a_i and b_i parameters are obtained to within .01 percent accuracy in 20 iterations giving an integral squared error of 0.0058152.

While not shown here, the matrix H is related to the Hessian matrix required in a second order Newton method [9, p 27]. Moreover, H becomes exactly equal to the Hessian as the values of the parameters approach an optimum. Hence this algorithm provides very rapid convergence in the neighborhood of an optimum.

VII. REFERENCES

1 R.N. McDonough and W.H. Huggins, "Best least-squares representation of signals by exponentials," IEEE Trans. on Automatic Control, Vol. AC-13, No. 4, pp 408-412; August 1968.

2 L.E. McBride, Jr., H.W. Schaefgen, and K. Steiglitz, "Time domain approximation by iterative methods, "IEEE Trans. Circuit Theory, Vol. CT-13, pp. 381-387, December 1966.

3 R.N. Chatterjee and M.M. Fahmy, "On time-domain approximation," Proc. of the 11th Midwest Sym. on Circuit Theory, pp. 288-298, May 1968.

4 R.D. Joseph and S.L. Huang, "Iterative method of real exponential approximation," Proc. of the 13th Midwest Sym. on Circuit Theory, pp. IV.2.1 - IV.2.6, May 1970.

5 G. Miller, "An iterative solution to the equations of Aigrain and Williams", IEEE Trans. Circuit Theory, Vol. CT-17, pp. 155-158, February 1970.

6 D.G. Luenberger, "Optimization by vector space methods," New York: John Wiley, 1969.

7 P.R. Aigrain and E.M. Williams, "Synthesis of n-reactance networks for desired transient response," J. Appl. Phys., Vol. 20, pp. 597-600, June 1949.

8 G. Miller, "Closed-form inversion of the gram matrix arising in certain least-squares problems," IEEE Trans. on Circuit Theory, Vol. CT-16, No. 2, pp. 237-240; May 1969.

9 E.J. Beltrami, "An Algorithmic Approach to Nonlinear Analysis and Optimization", Academic Press, 1970.

TEMPORAL PROPERTIES OF ENGINEERING SYSTEMS †

R. M. DESANTIS
Departement de Genie Electrique
Ecole Polytechnique
Montreal, Canada

W. A. PORTER
Department of Electrical Engineering
The University of Michigan
Ann Arbor, Michigan

ABSTRACT

The present article investigates the structure of engineering systems with respect to time related behavior. In a broad sense this study falls within the general context of causality theory. In particular it is consistent with the recent causality work of Porter [13], [14], and Saeks [15]. To distinguish this paper from the references it is noted that here linear and nonlinear systems as well as causal and noncausal are treated on an equal footing. In this regard a system model based on a new mathematical structure called Group Resolution Space is proposed. In this context various time related behavior classes of systems are considered and studied. They consist of the classes of causal, anticausal, and memoryless systems which have already been treated in the technical literature plus an additional class of systems called crosscausal. The main result of the paper can be stated as follows: every system in a Group Resolution Space is given by the sum of causal, anticausal, memoryless and crosscausal systems.

INTRODUCTION

For the description of problems related to physical systems the concept of time is frequently essential. In particular a physical system is called causal if present behavior of the system is only a function of the past. In this regard a widely accepted axiom is the principle of causality: all physical systems are causal.

When the concepts of time, and its associates - past, present and future -, are extended into the broader context of engineering systems they become abstract entities and, as a result, may preserve little in common with the original notion of physical "fourth dimension". In this case the terms real and nonreal time are often adopted.

In general, for engineering systems the assumption of causality is no longer valid. This is quite natural if it is considered that engineering systems need not be physical nor do they necessarily function in real time. Indeed in the engineering domain it is not unusual to be faced with problems

---

† This research was conducted at the University of Michigan and was in part supported by the NSF Grant GK1925

related to noncausal behavior. Classical filtering and estimation problems in communication theory, in general, do not have a causal solution ([2]). In stability theory the instability of a large class of linear time invariant systems can be explained in terms of absence of causality ([1],[18]). In optimal control theory noncausal controllers often occur in minimum effort, minimum fuel and optimal time problems ([12],[13]).

The above observations motivate the main thrust of the present development: to study definitely the modalities by which past present and future of the input affect the present of the output.

The principal factors to be considered are the following: i) the definition and the analysis of "basic" systems with a particular time related structure, ii) the investigation of the existence of the so called "canonical decomposition": can a general system be decomposed into the sum of basic systems? In a broad sense a study of this type falls into the context of classical causality theory ([6],[7],[11],[16],[17]). On the other hand, emphasis here is on systems which do not necessarily enjoy those properties on which much of the research on causality is based, namely: linearity, time invariance, continuity and most notable, causality (see [5] section 1.1).

It is not then surprising if the classical theory on causality presents limitations in this area. For instance, the technical literature has paid scant attention to the study of noncausal systems. Often this situation has been justified by the feeling that noncausal behavior can be treated as causal behavior in reverse time. This suits quite well studies confined to physical systems and to a certain extent to linear systems. However such an approach is not applicable to the type of problem considered here.

A need for the kind of study underlined above has been recently recognized in the work of Porter ([13],[14]) and Saeks ([15]). These references provide much of the ideas and inspiration of the present development. Particularly close in concept is [15]. Here Saeks defines and studies various classes of systems with "basic" time related properties. These classes consist of causal, anticausal, and memoryless systems. To categorize the time related structure of a linear system Saeks offers a canonical causality decomposition theorem. This theorem states that every linear system can be decomposed into the sum of causal, anticausal, and memoryless systems.

It may appear at this point as if the answer to the questions posed above had been exhausted. To show that this is not the case the following observations are in order: a) the mathematical framework adopted in [13], [15] limits the scope of the results to a restricted class of systems;

b) the statement of the canonical causality decomposition theorem, even in the limited form in which it has been presented, is not correct in general (see [5] section 4.7); c) the classes of causal, anticausal, and memoryless systems appear unable to account for the time related behavior of even simple nonlinear systems in communication and control theory.

The efforts of this study are directed to overcome some of the limitations and pitfalls indicated above. To this purpose a new mathematical framework called Group Resolution Space is adopted. In this setting the familiar classes of causal, anticausal, and memoryless systems appear together with a class of crosscausal systems. The concept of crosscausality is new and appears to be essential to extend the available theory of Porter and Saeks into a nonlinear systems context.

Within the proposed framework a correct and more general formulation of a canonical causality decomposition theorem can be given. A typical statement of this theorem goes as follows: every system can be decomposed into the addition of causal, anticausal, crosscausal, and memoryless systems.

MATHEMATICAL PRELIMINARIES

For brevity it is assumed that the reader is familiar with the notion of a linearly ordered set (p. 8 in [10]) and of a group (commutative or abelian (p. 32 in [10])).

Suppose that G is a group and ν a linearly ordered index set. The interest will be on $S[G,\nu] = \{x | x \text{ is a mapping } \nu \longrightarrow G\}$. In words if $x \in S[G,\nu]$ then to every $t \in \nu$, x associates to t a well defined element $x(t) \in G$. The space $S[G,\nu]$ comes equipped with a family of truncation operators P^t, P_t. These operators P^t, P_t act on $S[G,\nu]$ and are defined as follows: if x,y,z belong to $S[G,\nu]$ and $y = P^t x$, $z = P_t x$, then

$$y(s) = x(s) \text{ for } s \leq t,\ y(s) = \phi \text{ otherwise, and}$$
$$z(s) = x(s) \text{ for } s \geq t,\ z(s) = \phi \text{ otherwise,}$$

where the symbol ϕ indicates the null element of G. The space $S[G,\nu]$ together with the family of truncation operators will be called a Group Resolution Space (in short GRS).

In the sequel it will be found convenient to introduce an additional family of operators $dP(t): S[G,\nu] \longrightarrow S[G,\nu]$. This family is defined as follows: if x and y belong to $S[G,\nu]$ and $y = dP(t)x$, then $y(s) = x(s)$ if $s = t$ and $y(s) = \phi$ otherwise. Clearly $dP(t) = P^t P_t = P_t P^t$. For later use the notations $\overline{P}^t$ and $\overline{P}_t$ will indicate the following operators

$$\overline{P}^t = P^t - dP(t) \text{ and } \overline{P}_t = P_t - dP(t) .$$

Occasionally $dP(t)x$, $\overline{P}^t x$ and $\overline{P}_t x$ are referred to respectively as present, past and future of x at time t.

A GRS can be easily given a group structure. To this purpose the addition of two elements is defined as follows: if $x,y,z \in S[G,\nu]$ then z is the addition of x and y, $z = x+y$, when $z(s) = x(s) + y(s)$ for all $s \in \nu$. Similarly z is the difference between x and y, $z = x-y$ if $y(s) + x(s) = z(s)$. For later use it is observed that by virture of these definitions, any element $x \in S[G,\nu]$ can be represented as follows: $x = \sum_{t \in \nu} dP(t)x$. Similarly $P^t x$ and $P_t x$ can be represented by

$$P^t x = \sum_{s \le t} dP(s)x, \quad P_t x = \sum_{s \ge t} dP(s)x.$$

Consider now operators T, T', T" all mapping $S[G,\nu] \longrightarrow S[G,\nu]$. The operator T is the addition of T' and T", $T = T' + T''$ if for every $x \in S[G,\nu]$ the following relation holds: $Tx = T'x + T''x$. T is the difference between T' and T", $T = T' - T''$, if $Tx = T'x - T''x$. T is the composition of T' and T", $T = T'T''$ is for every $x \in S[G,\nu]$ the following relation holds: $Tx = T'(T''x)$.

An operator T on a GRS is unbiased if $T[\phi] = \phi$ where ϕ indicates the element x in $S[G,\nu]$ such that $x(s) = \phi$ (the null element in G) for all $s \in \nu$. In the sequel only unbiased operators will be considered. This is not a serious limitation. If T is not unbiased, the development will still be valid for the operator $T - T[\phi]$.

TIME BEHAVIOR RELATED CONCEPTS

In this section some basic concepts related to causality are introduced and discussed. The notation T is used to indicate an operator (unbiased) on $S[G,\nu]$ and x and y indicate two elements in $S[G,\nu]$.

Definition 1. T is causal (anticausal) if

$$P^t Tx = P^t Ty \text{ when } P^t x = P^t y \quad (P_t Tx = P_t Ty \text{ when } P_t x = P_t y)$$

Definition 2. T is memoryless if it is simultaneously causal and anticausal.

Definition 3. T is strongly causal (strongly anticausal) if T is causal (anticausal) and

$$TdP(t)\, x = \phi$$

Definition 4. T is strictly causal (strictly anticausal)

$$P^t Tx = P^t Ty \text{ when } \overline{P}^t x = \overline{P}^t y \ (P_t Tx = P_t Ty \text{ when } \overline{P}_t x = \overline{P}_t y).$$

Definition 5. T is crosscausal if:

$$P_t Tx = \phi \text{ when } P_t x = \phi, \ P^t Tx = \phi \text{ when } P^t x = \phi, \text{ and } TdP(t)x = \phi.$$

Definition 6. T is strongly crosscausal if

$$P_tTx = \phi \text{ when } \overline{P}_tx = \phi, \text{ and } P^tTx = \phi \text{ when } \overline{P}^tx = \phi.$$

The concepts of causal, anticausal, and memoryless systems are in formal agreement with those proposed by Porter [13], [14] and Saeks [15]. When the notion of integral is identified with the notion of sum, this agreement extends also to the concepts of strictly causal and strictly anticausal systems.

The concepts of crosscausality and strong causality find no correspondence in previous work of other authors. It will become clear with the unfolding of the present development, that these concepts are essential for the study of the causality structure of nonlinear systems. In [5] it is shown that in the case of additive systems the concept of crosscausality vanishes and strong causality coincides with strict causality. This explains in part the absence of these concepts in the classical literature.

In the sequel it will be helpful to refer to the causality related concepts through an alphabetic code. For this purpose the following short hand notations will be adopted.

A = anticausal
C = anticausal
M = memoryless
$\overline{A}$ = strictly anticausal
$\overline{C}$ = strictly causal
X = crosscausal
$\underline{X}$ = strongly crosscausal
$\underline{A}$ = strongly anticausal
$\underline{C}$ = strongly crosscausal

CAUSALITY RELATED PROPERTIES

In this section various properties associated with those systems which enjoy a special structure with respect to causality are presented. As in the previous section the notation T will be used to indicate a system defined on a GRS. For simplicity the mathematical proofs have been omitted and the main emphasis is given to engineering interpretation. The reader interested in a detailed account of the proofs is referred to [5], Chapter 2.

Proposition 1. The following statements are equivalent:

a) T is causal
b) $P^tT = P^tTP^t$ for all $t \in \nu$
c) T has the following representation $T = \sum_{s\in\nu} dP(s)TP^s$.

Proposition 2. The following statements are equivalent

a) T is memoryless
b) $P^tT = P^tTP^t$ and $P_tT = P_tTP_t$
c) $T = \sum_{s\in\nu} dP(s)TdP(s)$.

Proposition 3. The following statements are equivalent:

a) T is strongly causal

b) $P^tT = P^tTP^t$ and $dP(t)\ TdP(t) = \phi$

c) T is causal and $\sum_{s \in \nu} dP(s)\ TdP(s) = \phi$.

Proposition 4. The following statements are equivalent:

a) T is strictly causal

b) $P^tT = P^tT\overline{P}^t$

c) $T = \sum_{s \in \nu} dP(s)T\overline{P}^s$.

Proposition 5. The following statements are equivalent:

a) T is crosscausal

b) $TP^t = P^tTP^t$, $TP_t = P_tTP_t$ and $TdP(t) = \phi$

c) $\sum_{s \in \nu} dP(s)T\overline{P}^s = \sum_{s \in \nu} dP(s)\ T\overline{P}_s = \sum_{s \in \nu} dP(s)\ TdP(s) = \phi$.

Proposition 6. The following statements are equivalent:

a) T is strongly crosscausal

b) $TP^t = \overline{P}^t\ TP^t$ and $TP_t = \overline{P}_tTP_t$

c) $\sum_{s \in \nu} dP(s)TP^s = \sum_{s \in \nu} dP(s)TP_s = \phi$.

In most nontrivial applications basic subsystems are combined through the operations of addition and composition. In this regard the following results apply.

Proposition 7. Suppose that the operators T and T' are α for some $\alpha \epsilon \{A,C,M,X,\underline{A},\underline{C},\overline{A},\overline{C},\underline{X}\}$. Then T + T' and TT' are also α.

Proposition 8. If T is memoryless and T' is $\alpha \epsilon \{A,C,M,X,\underline{A},\underline{C},\overline{A},\overline{C},\underline{X}\}$ then T'T and TT' are also α.

Proposition 9. If T is strongly (strictly) causal and T' is causal then TT' and T'T are also strongly (strictly) causal.

Proposition 10. If T is strongly crosscausal and T' is crosscausal then TT' and T'T are strongly crosscausal.

To conclude this review of causality properties observe that the various causality classes under consideration are, in general, overlapping. For instance, a memoryless systems is also causal and anticausal. Similarly a system which is crosscausal can, at the same time, be causal or anticausal. The next result indicates, however, that the classes of strongly causal, strongly anticausal, strongly crosscausal and memoryless systems are disjoint.

Proposition 11. If the system T is simultaneously α_1 and α_2, with α_1, $\alpha_2 \in \{\underline{A}, \underline{C}, M, \underline{X}\}$ then either $\alpha_1 = \alpha_2$ or T is the null operator.

CANONICAL TIME RELATED DECOMPOSITION

The development in the previous sections allows to gain a certain insight into the causality structure of those systems which are basic, that is have the property to be $\alpha \in \{A,C,M,X,\underline{A},\underline{C},\underline{X},\overline{A},\overline{C}\}$. The purpose of this section is to develope a similar insight in the case of more general systems.

To help motivate the approach, observe that the insight into the causality structure of basic systems can be readily used to characterize the causality structure of those systems which are given by the addition of basic systems. The natural question then arises whether it is possible to represent a general system as the addition of basic systems. If this representation does exist it also is of interest to know whether it is unique.

The above question has already been considered in the technical literature and goes under the name of canonical causality decomposition question (Saeks [15]). However, while the systems considered by Saeks are linear and defined on a Hilbert space, the systems in this development are neither linear nor are they defined in a linear space. Thus, some further elaboration is needed in order to obtain a satisfactory answer.

The matter of uniquess is not available in general. Suppose, for example, that $T = T_A + T_C$. Then we can also write

$$T = T'_C + T_A, \text{ where } T'_A = T_A + T_M,\ T'_C = T_C + T_M \text{ and}$$

T_M is any memoryless operator. This implies that the decomposition of T is not unique.

A moment of reflection shows that the above argument is successful because the various classes of basic systems are overlapping. On the other hand we have seen in proposition 4.11 that the classes of systems $\alpha \in \{\underline{A},\underline{C},M,\underline{X}\}$ are disjoint. From here it appears reasonable to modify the causality decomposition question with the following more specific terminology: "is it possible to decompose a general system into the sum of systems with are $\alpha \in \{\underline{A},\underline{C},M,\underline{X}\}$? If a decomposition of this type exists it will be called a canonical causality decomposition.

Proposition 1. If a canonical causality decomposition exists then it is unique.

In general the question of existence of a canonical decomposition can be viewed in terms of the existence of a special set of mappings on the space of operators. This point of view offers some technical advantages and it is convenient to clarify it.

To being with, let $\$$ indicate the space of all operators T on a GRS, and $\bar{I}$ an operator on $\$$. To stress that $\bar{I}$ maps a space of operators into itself, $\bar{I}$ is called a transformator [9]. A transformator $\bar{I}$ will be indicated by $\bar{I}_\alpha$, $\alpha \in \{\underline{A},\underline{C},M,\underline{X}\}$ if for each $T \in \$$ we have that $\bar{I}_\alpha [T]$ is an α operator. For example the set $\underline{X}$ is invariant under $\bar{I}_X$. Occasionally we will also use the notation I to indicate the identity transformator, that is the transformator with the property that $I[T] = T$ for every T in $\$$.

The following proposition indicates the equivalence between existence of a canonical decomposition and existence of a special set of transformators.

Proposition 2. A canonical causality decomposition always exists if and only if there exists a set of transformators $\bar{I}_\alpha$, $\alpha\in\{\underline{A},\underline{C},M,\underline{X}\} = \mathcal{A}$ such that the following properties are satisfied:

pi) $\bar{I}_\alpha$ is defined on $\$$

pii) For each pair $(\alpha,\gamma) \in\mathcal{A}^2$ we have that

$$\bar{I}_\alpha \bar{I}_\gamma = \emptyset \text{ if } \alpha \neq \gamma \text{ and } \bar{I}_\alpha \bar{I}_\gamma = \bar{T}_\alpha \text{ if } \alpha = \gamma$$

piii) $\sum\limits_{\alpha\in\mathcal{A}} \bar{I}_\alpha = I.$

With the above results it is now easy to state and prove the following proposition. The proof is included because it is representative of the type of techniques involved in the overall development.

Proposition 3. Every system on a GRS can be uniquely decomposed into the sum of strongly causal, strongly anticausal, strongly crosscausal and memoryless components.

Proof. Suppose that T is an operator on $S[G,\nu]$ and define a transformator $\bar{I}_M$ according to the following rule:

$$\bar{I}_M[T] = \sum_{s\in\nu} dP(s)TdP(s).$$

From Proposition 4.2, $\bar{I}_M[T]$ is a memoryless operator. Define the transformators $\bar{I}_C$, $\bar{I}_A$ as follows

$$\bar{I}_C[T] = \sum_{s\in\nu} dP(s)\,.[T-\bar{I}_M[T]]P^s$$

$$\bar{I}_A[T] = \sum_{s\in\nu} dP(s)\ [T-\bar{I}_M[T]]P_s\ .$$

From Proposition 4.1 and its dual, $I_{\underline{C}}[T]$ and $I_{\underline{A}}[T]$ are respectively strongly causal and strongly anticausal. Finally, consider the transformator $\bar{I}_X$ defined by the following expression

$$\bar{I}_{\underline{X}}[T] = T - \bar{I}_M[T] - \bar{I}_{\underline{C}}[T] - \bar{I}_{\underline{A}}[T]$$

Observe that the operator $\bar{I}_X[T]$ satisfies the relations in proposition 4.6. Hence, it is strongly crosscausal. At this point verify that the above set of tranformators $\bar{I}_M$, $\bar{I}_C$, $\bar{I}_A$, and $\bar{I}_X$ satisfies Pi-ii-iii) in proposition 2. It can be concluded that $\bar{T}$ has a causality canonical decomposition.

Combining Propositions 1 and 3, it is now possible to state the following *Canonical Causality Decomposition Theorem*.

Theorem 1. Every system on a GRS has a canonical causality decomposition and this decomposition is unique.

APPLICATIONS

Concepts and ideas developed in the previous sections play an important role in various engineering fields. As a consequence potential applications are numerous. A proper characterization of their extent and scope goes beyond the objectives of the present paper. For completeness however, a specific application with both conceptual and practical overtones will be briefly discussed.

This application can be found in the work reported in [3] and [4]. In these references concepts, ideas, and results from the present development are extended in a natural way to systems defined in Hilbert space. They are then applied to a well studied feedback stability problem. From here results from the theory of linear non-self-adjoint operators, ([9]), are given a meaningful physical interpretation in terms of modern stability theory [3]. Moreover new results of practical interest are also generated. This is amply illustrated in [4]. Typical among these results is for example the following: a sufficient condition for a closed loop feedback system to be stable is that the open loop system be weakly additive bounded continuous and strictly causal.

CONCLUSIONS

The contribution of this paper is characterized by two main aspects: formulation and theoretical results.

The formulation consists of the development of a fertile mathematical setting to study some basic causality questions. It contains two novel ideas: the use of the structure of Group Resolution Spaces and the introduction of the class of crosscausal operators. These two ideas are exploited to discuss with a certain mathematical rigor various important properties connected to system time-related behavior. These properties are emphasized by a number of a propositions and theorems in sections 4-5 and 6.

The most import theoretical result consists of theorem 5.1. This result give a general canonical decomposition theorem for nonlinear operators on nonlinear spaces. It can be stated as follows:

every system in a Group Resolution Space can be decomposed into the sum of strongly causal, strongly anticausal, memoryless and strongly crosscausal systems.

ACKNOWLEDGEMENT

The authors wish to extend their thanks to Professor Richard Saeks for helpful discussions.

REFERENCES

1. M. Damborg, "Stability of the Basic Nonlinear Feedback System", Tech Rep. 37, Systems Engineering Laboratory, University of Michigan, Ann Arbor, 1969.
2. W. B. Davenport and W. L. Root, An Introduction to the Theory of Random Signals and Noise, (McGraw-Hill, New York, 1958).
3. R. M. DeSantis, "On Some Connections Between Causality and Stability", preprints Midwest Symposium on Circuit Theory, University of Denver, May 1971.
4. R. M. DeSantis, "On Strict Causality and Stability", submitted to the IEEE Trans. on Automatic Control, April 8, 1971.
5. R. M. DeSantis, "Causality Structure of Engineering Systems", Ph.D. dissertation, University of Michigan, Ann Arbor, Dec. 1971 (in preparation).
6. P. L. Falb and M. I. Freedman, "A Generalized Transform Theory for Causal Operators", SIAM J. Control, vol. 7, no. 3, August 1969.
7. Y. Foures and I. E. Segal, "Causality and Analycity", Trans. Amer. Math. Soc.,78 (1955) pp. 385-405.
8. A. Gersho, "Nonlinear Systems with a Restricted Additivity Property", IEE Trans. on Circuit Theory, vol, CT-16, no. 2, (1969) pp.150-154.
9. I.Z. Gohberg and M. Krein, Theory of Volterra Operators in Hilbert Space and Applications, Amer. Math.Soc., vol. 24, (translation) 1970.
10. E.Hewitt and K.Stromberg, Real and Abstract Analysis, Springer-Verlag, 1969,
11. R. E. Paley and N. Wiener, Fourier Transform in the Complex Domain, Amer. Math. Soc., Colloquium Publ., vol. 19, 1934.
12. W.A. Porter, Modern Foundations of Systems Engineering, (MacMillan, New York, 1966).
13. W.A. Porter and C.L. Zahm, "Basic Concepts in System Theory", Tech.Rep. SEL 44, The University of Michigan, Ann Arbor, 1969.
14. W.A. Porter, "Some Circuit Theory Concepts Revisited", Int. J. Control, vol. 12, no. 3, (1970), pp. 433-488.
15. R.Saeks, "Causality in Hilbert Space", SIAM Review, vol. 12, (1970) pp. 357-383.
16. J.S. Toll, "Causality and the Dispersion Relation: Logical Foundations", The Physical Review, vol. 104, no. 6, December 1956.
17. E.P. Wigner, Dispersion Relations and Their Connection with Causality, (Academic Press - New York and London, 1964).
18. J.C. Willems, "Stability, Instability, Invertibility and Causality", SIAM J. on Control, vol. 7, 1969, pp. 645-671.

URBAN HEALTH CARE SIMULATION: AN ECOLOGICAL APPROACH TO SYSTEM ANALYSIS AND EVALUATION

J. MILSUM
Biomedical Engineering
McGill University
Montreal, Quebec

R. STROHMAIER
Health Systems Group, Resource Science Centre
The University of British Columbia
Vancouver, B. C.

D. UYENO
Management Science
The University of British Columbia
Vancouver, B. C.

I. VERTINSKY
Resource Science Centre
The University of British Columbia
Vancouver, B. C.
and
Epidemiology
Tulane School of Public Health
and Tropical Medicine
New Orleans, Louisiana

H. WILL
Accounting and Management Information Systems
The University of British Columbia
Vancouver, B. C.

ABSTRACT

The Health Care Simulation Model is part of a microcosmic model of the Greater Vancouver Region. This Health sub-model is characterized by interactive capabilities in order to provide operational indicators for the impact of changes in other sub-models. Morbidity prevalence levels in the Health sub-model are modified by outputs from the Population, Pollution and Transportation sub-models. These are modified by outputs from the Human Ecology sub-model to generate demands for health services, which counteract environmental and social stress. The health resources and demands are matched to measure the extent of these remedial actions.

I. INTRODUCTION

The total North American bill for health expenditures has been increasing at annual rates of between 10 and 15% for a number of years. If this were to continue, in 3 to 5 years we would be devoting about 10% of GNP to health care, up from 5 to 7% presently. In 9 to 14 years this could rise by virtue of the compounding effect of annual rates to 25% of GNP. The health care system is but one of the sectors to which the consumer can allocate his resources. Most would readily agree that the trend in health care costs must be dampened without significant detrimental effects on health status.

An ecological perspective of the health care system recognizes that there are environmental factors such as sanitation level, quality of air, and life styles which influence morbidity and mortality patterns. Consequently, the identification of cost-effective strategies of health care necessitates the consideration of activities which traditionally may not be regarded as being within the domain of health care systems evaluation. To this end, we have attempted to develop an urban health care simulation model within the context of an overall regional simulation.

The health care system is structurally defined in terms of human and material health care resources; e.g. physicians, nurses and beds. Operationally, these components interact to treat incidences of morbidity according to the state of the art. The conditions of life in a society determine both its demand for health care resources and its supply of these resources for health care delivery. In order to analyze the magnitude of future health care demand, it is necessary to determine the cause-effect relationships of environmental conditions on individual and social health.

Since it is impossible to study this problem in its total proportions, we decided to simulate urban health care in a microcosmic setting. As part of a larger simulation project called Inter-Institutional Policy Simulator (IIPS) of the Greater Vancouver Region, we are trying to analyze and evaluate the region's health care system.

> By late spring [1970], the City of Vancouver, the Greater Vancouver Regional District, the University of British Columbia and the Ford Foundation had all agreed to a five-year commitment to develop a large-scale simulation model for the Greater Vancouver Region, on which a variety of policies might be tested in a simulated world, so that unexpected and deleterious consequences of these policies might be ascertained *ex ante*.[2]

The health care simulation model can therefore be considered as one module of the larger simulation project and is designed to interact with other modules such as the population, human ecology, pollution and transportation models as shown in Figure 1.

The decision on breakdown of the urban system into subsystems for simulation purposes was an important one. Typically, a practical criterion is that each sub-system should concern an aspect which merits study in its own right and which is not cross-connected to the other subsystems through such a dense network of variables as to render the idea of separation meaningless. There is always the danger that the very selection of such subsystems could force the modeling, the policy proposals, and the solutions ultimately adopted into certain preset modes. Thus one expressed purpose of the modeling, that of being able to propose and test radically new solutions could be prevented *ab initio*, i.e. the modeling could necessarily be a failure by restricting the conception of the system and its testing to limited and probably unsatisfactory ways.

In the present study this problem is well illustrated by the health care system. Thus we do not consider the subsystem of medicine alone, as provided through the hospitals, physicians' practices, and pharmacies. Rather we consider the health care system with its implied attention to various preventative measures for maintaining well human beings.[1] Implied is the concept that some performance measures must be sought so that the

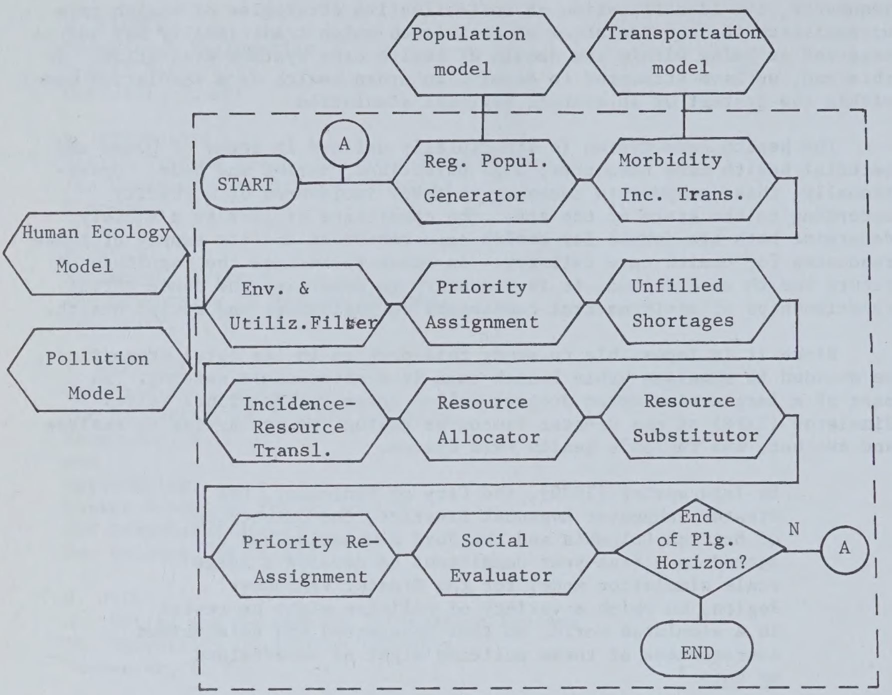

Population model
Transportation model
A
START
Reg. Popul. Generator
Morbidity Inc. Trans.
Human Ecology Model
Pollution Model
Env. & Utiliz. Filter
Priority Assignment
Unfilled Shortages
Incidence-Resource Transl.
Resource Allocator
Resource Substitutor
Priority Re-Assignment
Social Evaluator
End of Plg. Horizon?
N
A
END

Figure 1: Urban Health Care Simulation Model

cost:effectiveness levels of systems with greatly different priorities in policy can be at least qualitatively ordered with some confidence.

We contend that the health system, through its patterns of morbidity, mortality and disability provides possibly a simple indicator of a meaningful quality of life index. This contention is supported by recent research that intimately ties the abilities to cope with environmental and social stresses with the visible output of morbidity. Morbidity data can subsequently be translated into health care demands and related to health care supply, performance, satisfaction and efficiency. To simulate these causes and effects for the Greater Vancouver Region is the major purpose of our Urban Health Care Simulation Model.

II. HEALTH CARE SIMULATION MODEL

Our model, while centering on the provision of remedial and preventative services, takes into consideration the links to those submodels which generate stressors (human ecology, pollution, transportation, etc.) and projects their effects (when knowledge of cause and effect is available) on visible patterns of morbidity, mortality and disability. Hence, through the model, stressors and the coping abilities(of which health care systems are often only a final remedial step) are matched to provide output which may serve as an indicator of quality of life. This present notion of quality of life is independent of any absolute notions of a "good life", and centers instead on a paradigm of optimal adaptation from which any deviation is manifested by lack of health and by death.

At the moment, the model is concerned primarily with the provision of health care. While environmental and social impact are expressed through readjustment of "needs" for health services, we hope to broaden the perspective to provide more emphasis on the ecological aspects.

A. Outline Description of Model

The model presently comprises four major component processes:

(i) Disease Generator
(ii) Priority Streaming
(iii) Resource Allocation for Treatment
(iv) Feedback of System Shortages and Estimation of System Shortages

As noted above, the model should ideally provide estimates for Social Indicators of the Benefits, Satisfactions and Dissatisfactions associated with the Health Care System. Through the models of other subgroups, in particular Human Ecology, the political process should be modelled to relate the indicators to expenditures on those processes which have direct or indirect effect on the Health Care System's efficiency and capacity. In particular these expenditures are for capital equipment, health-related education programs, and environmental and pollution programs. The feedback of Social Indicators should influence the simulated expenditure on these programs and the appropriate allocation. Clearly this is impossible at present. This feedback must constitute one of the important interventions for our model.

Disease Generation. The population is classified by age and sex, but not by socio-economic condition or geographic subareas in estimating the "natural" prevalence of disease. Diseases and services are aggregated into 27 groups. In addition, casualty figures are provided by the Transportation and Industrial groups. Exogenous data from other groups (notably, Transportation and Pollution) enter somewhat obliquely here in that they may alter the "natural" prevalence rates. Such data are related to the particular diseases they affect by the use of dominating factors. These dominating factors are expressed as coefficients which modify incidence rates. As an example, Respiratory Diseases are affected by air pollution.

Since the extent to which the population perceives the existence of a disease, and the extent to which people may then seek treatment, both vary according to social awareness and demand, a further filtering process is now applied on the various disease rates. Specifically, these utilization factors are classified by disease category. The output of the disease generation process is thus a set of prevalence rates classified by disease, age and sex.

Priority Streaming. This section of the model recognizes that the system inevitably applies some priorities in its evaluation of what needs are most urgent. In brief, each disease category is allocated a basic priority level. For patients who do not get accepted for treatment in one time period, a higher priority is assigned in the next time period. This is, of course, exemplified by the patient for elective surgery who may be delayed as opposed to the accident victim who must be treated immediately.

Resource Allocation for Treatment. The resources are classified by physician specialty (physician, surgeon, pediatrician, nursing hours and bed days. The patients with highest priority are then matched with their resource needs, and decremented values of the resources are computed. This resource allocation is continued until a shortage in some resource is detected. Substitution of resources may be arranged in some specific allowable cases. For example, some classes of MD's can substitute for each other, and MD's can even substitute for nurses.

Feedback of System Shortages and Estimation of System Shortages. The resource shortages and the numbers of untreated cases are then computed for the given time unit. The number of cases treated by disease category is then determined. These data on treated and untreated cases provide the basis for the generation of estimates for meaningful Social Indicators for system performance.

B. Features of the Current Model

1. The creation of a normative model of the health care system as well as a model of newly proposed health care patterns. Examples:

a) With slight modification of demand matching, the simulation may describe various degrees of centralization in the provision of health care. The original model proposes a central resource-demand matching strategy, but it can be modified to provide resource matching on functional lines (e.g. specialized hospitals for cancer or T.B.) or on geographical lines (integration of hospital districts, joint admission bureaus, etc.).

b) Patterns of resource utilization such as delegation of functions can be modified by changing the rules on the resolution of shortages in particular categories of health resources.

2. The effect of altered priority systems.

3. The effect of changing resource requirements due to new technology or altered organizational patterns.

4. Prediction of the occurrence of major bottlenecks in the current system. Consequently a base could be provided to evaluate directions for reform.

5. A flexible structure for the incorporation of submodels (in particular the submodel of health care utilization and the impact of environmental variables). Morbidity-mortality patterns are set at present through fixed parameters for each run. By parametric variation of the dominating factors, one can incorporate environmental factors such as noise, water and air pollution, deterioration of sanitation levels and changes in health behaviour such as nutrition behaviour, drug use, and intensity of life crises.

C. Applicability of the Model

1. Time Determination and Health Care Planning: We have adopted a yearly feedback model. This time frame is especially amenable to current

practices of centralized decision-making in the region. In the Vancouver region, health care capital budgeting is done on a multi-year basis by a Regional Hospital District Board.

2. Determination of Geographic Areas: Long range planning and capital budgeting are done at present on the basis of hospital districts which have geographic responsibility. The Greater Vancouver Hospital District accounts for slightly more than half the people in B.C.

D. A Mechanism for System Evaluation

To evaluate the system's performance under the various experimental conditions we have explored a new approach in the weighting of shortage costs and in obtaining an indicator of their possible political and social impact. Holmes and his colleagues,[8] have constructed a quantitative scale for rating the seriousness of the most common and representative illnesses. This Seriousness of Illness Rating Scale (SIRS) has been shown by them to have both satisfactory reproducibility[9] and agreement between the scaling by medical and non-medical people. It is based on personal perceptions of the seriousness of an illness, given just one normalizing value, which was specified as 500 for Peptic Ulcer. (ranking 89 amoung the conditions in the survey).

Presumably the maximum benefits are achieved when unmet demands weighted by the seriousness scale and the promptness of delays in treatment are minimized. We are at present unable to impute a factor for situations in which patients receive treatment but are not restored to a normal state of health. Achieved benefits are measured as the summation of treated cases, suitably weighted by the SIRS scale and by the promptness of treatment.

The existence of the SIRS scales enables the model to operate either in terms of the preferences established by physicians or those established by consumers. In the current model, the priority streaming is dependent on the medical preferences whereas the evaluation of social impact is evaluated on the preferences of the non-medical group.

We should also note that many other illnesses are known to be psychosomatically related to stress level, for example High Blood Pressure (91, 520), Peptic Ulcer (89, 500), Migraine (56, 242) and Eczema (45, 204). Indeed Holmes and his colleagues[3] have established a quantitative scale (Social Readjustment Rating Scale) for the perceived seriousness of various "life events" such as marriage, divorce, unemployment, etc. By accumulating these "Life Change Units" over a period of time (typically two years) a probability of illness onset can be assigned for the next short period (typically six months). From such work there begins to emerge an exciting conjecture that individual health may serve as a causally-meaningful indicator for social health or vice versa. That is, the quality of life in the larger social system of which the health care system is a part depends on the health of individuals as much as on the health level of society. Hence in our sub-system an adequate performance measure must ultimately be able to incorporate the decrease of illness prevalence rates as a "good". Thus the optimum in achieved benefits would be partially dependent on maintaining the lowest possible prevalence rates channeled through the medical care system.

III. EXPERIMENTS

To analyze and evaluate the behaviour of the regional health model, four scenarios have been simulated. These are based on conjectures about the effect of selected environmental factors such as traffic accidents and air pollution on health care under two alternative population growth patterns. The domination factors for those effects are varied parametrically. The experiments may also be considered as a test of the interface of the health care model with the Pollution and Transportation models which are in development.

The experimental conditions chosen to demonstrate the capabilities of the model include the following configurations:

1. Normal population growth with no change in environmental conditions. This experimental run aims at establishing a basic pattern to which experiments with varying environmental and demographic conditions can be compared.

2. Normal population growth with environmental deterioration. This experiment represents a 10% parametric adjustment in morbidity related to traffic accidents, and environmental pollution levels.

3. A "zero population growth" (ZPG) intervention with no change in environmental conditions.

4. A ZPG intervention with similar deterioration in environmental conditions as specified in experiment 2.

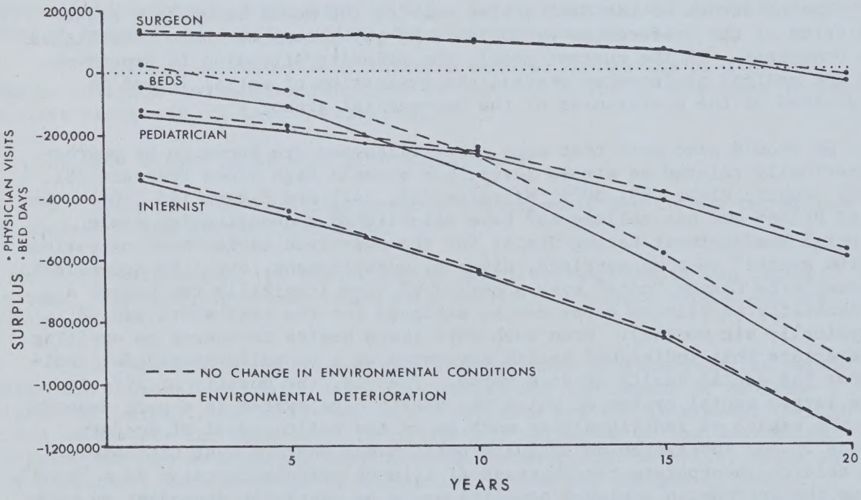


Figure 2: Selected Health Resource Surpluses Under Two Environmental Conditions with Normal Population Growth

Resources were incremented annually in experiments 1 and 2 by 2%, while no resource growth was permitted under a ZPG intervention.

Under normal population growth the regional population increased from 1,116,703 in the first simulated year to 2,661,826 in the 20th year.

ZPG was interpreted in the experiment as a limitation of new births to those women with less than 2 children. Under this policy the population reached the level of 1,277,183 in the 20th year.

The simulation provided data on annual morbidity patterns for 27 diagnostic or service classes. These are reflected in the resource surpluses or shortages for 12 resource categories. They are general practitioners, internists, pediatricians, surgeons, obstetricians, anesthesiologists, opthalmologists, otolarngologists, orthopedic surgeon, urologist, nurses, bed days. The surpluses or shortages for four selected resources are represented graphically for each experimental condition in Figures 2 and 3. These are surgeons, internists, pediatricians and hospital beds. It should be noted that these are computed before substitution of resources is performed by a delegation routine.

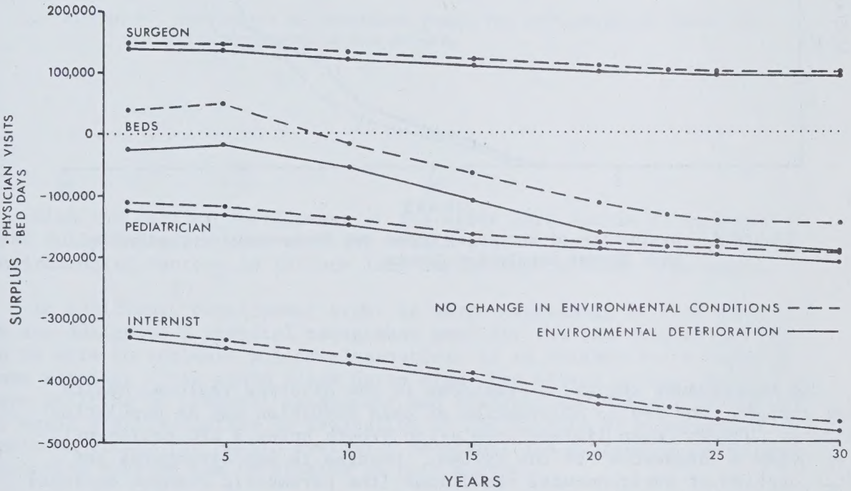


Figure 3: Selected Health Resource Surpluses Under Two Environmental Conditions with Zero Population Growth

In the experiments described in this paper SIRS weights were applied to untreated episodes of diseases to obtain an indicator of the system's performance from the non-medical point of view. The graphs presented in Figures 4 and 5 represent the system's performance under the four experimental conditions. When the index of untreated cases is zero, the system satisfies all demands. To ensure comparability, similar resource growth ratios in all the experiments were maintained. In computing the number of untreated cases the model employed the delegation routine permitting some reasonable substitutions between health resources.

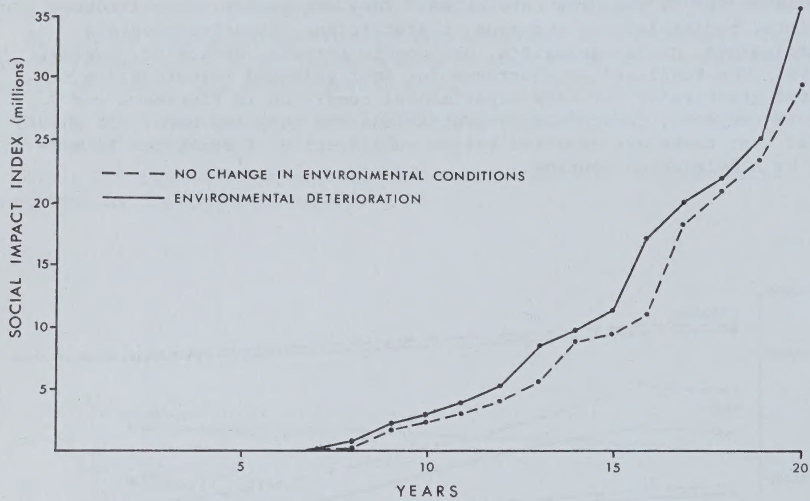


Figure 4: Evaluation of Shortages Under Two Environmental Conditions with Normal Population Growth

The experiments suggest a breakdown in the existing regional health care system. The key to alleviation of this situation may be population control. However even limited population growth under a ZPG policy will only delay a "breakdown" of the system. Changes in age structures and deterioration of environmental conditions [the parametric changes employed for environmental conditions in the experiment are "optimistic"] lead even under a ZPG intervention to emerging shortages before the 20th simulated year.

IV. EXTENSIONS

As mentioned in section II, part D, further extensions of this model are planned with respect to additional health care services and delegation alternatives that are available to the professions within each kind of service. These include the dental care services which were left out of the current design, delegation opportunities within dental health care, and an evaluation of concerted preventative care programs in such areas as dental health.

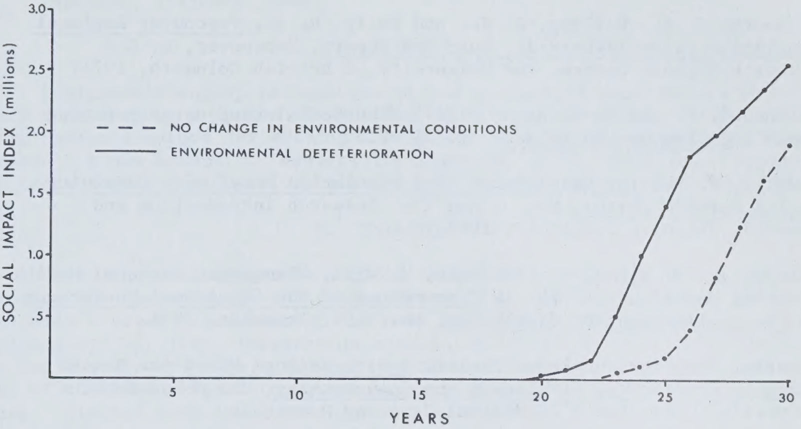


Figure 5: Evaluation of Shortages Under Two Environmental Conditions with Zero Population Growth

With the further development of the other IIPS models it is hoped to more fully integrate them with the health systems model by extending the environmental factors to include life styles and relevant life events.

An additional development might be worth mentioning in this context. We are designing a hospital management game for the same region and hope to be able to evaluate policy alternatives in an interactive simulation - game setting. This would allow us to study the effect of various health care policies on hospital management and hospital development in the region. It would also facilitate an evaluation of the effects of behavioural patterns within the health systems. Moreover, the influences of various incentive schemes can be studies in this manner.

V. CONCLUSIONS

The above experiments were merely illustrative of ecologically relevant parametric changes of the urban health care simulation. Elsewhere we have shown how the model can be used to evaluate alternative growth patterns of health care resources.[6]

The quality of life is of major interest in this regard. Urban health care simulation is an attempt to analyze and to evaluate it as comprehensively as possible, by applying general systems theory, management information systems design and systems management within a regional ecology.

REFERENCES

1. "Conceptual Problems in Developing an Index of Health", PHS 1000-2-17, Washington, May, 1966.

2. Goldberg, M. A., Holling, G. S., and Kelly, R. F., Vancouver Regional Simulation Study 1970-1971: Year End Report, Vancouver, B. C.: Resource Science Centre, The University of British Columbia, 1971.

3. Holmes, T. H. and R. H. Rahe, (1967) "The Social Readjustment Rating Scale", J. Psychosomatic Res. 11: 213-218.

4. Kennedy, F. D., The Development of a Simulation Model of a Community Health Service System, Vol. I and II. Research Introduction and Summary, Triangle Institute, 1968.

5. Milsum, J., D. Uyeno, I. Vertinsky, H. Will, "Vancouver Regional Health Planning Model", to appear in Proceedings of the 5th Annual Conference on the Applications of Simulation, New York, December, 1971.

6. Navarro, Vincente and Roger Parker, A Mathematical Model for Health Planning Prediction, Simulation and Goal-seeking, The Johns Hopkins University, Department of Medical Care and Hospitals.

7. Wong, A.K.C. and Tung, Au, "A Dynamic Model for Providing Patient Care in Hospitals", Proceedings of the 4th Conference on the Applications of Simulation, New York, 1970.

8. Wyler, A. R., M. Masuda, and T. H. Holmes, (1968) "Seriousness of Illness Rating Scale", J. Psychosomatic Res. 11: 363-374.

9. _________ (1970), "Seriousness of Illness Rating Scale", J. Psychosomatic Res., 14: 59-64.

A COMPUTER APPROACH FOR ANALYSIS OF NONLINEAR CIRCUITS[1]

D. A. R. ZEIN AND S. L. HAKIMI
Department of Electrical Engineering
Northwestern University
Evanston, Illinois 60001

ABSTRACT

Liniger-Willoughby integration method is used to describe a circuit at each time step by a system of nonlinear algebraic equations. A controlled-step Newton's method for solving the system of algebraic equations is presented. A new method for sensitivity computation is introduced. Finally, optimization of nonlinear circuits is briefly described. Examples are given.

1. INTRODUCTION

This paper describes the various facets of a general purpose computer program for the analysis and optimization of nonlinear circuits. We start out with a brief discussion of the main features of the Liniger-Willoughby [1], (LW), integration method. Then the LW integration method, in conjunction with the node equations of the circuit, is used to generate a set of nonlinear algebraic equations that describe the circuit at every time step. Similar technique was suggested by Calahan [2] in connection with Gear's [3] integration method. A brief theoretical discussion of a controlled-step Newton's method [4] for solving a set of nonlinear algebriac equations is presented. We also discuss a new method for computing the sensitivity matrix $\frac{\partial X}{\partial P}$, where $X = (x_1, x_2, \ldots, x_n)^t$, and $P = (p_1, p_2, \ldots, p_k)^t$, are the solution vector and a circuit parameter vector, respectively. The main features of a general purpose analysis-optimization program is discussed and examples, which demonstrate the flexibility and capability of our method, are given.

2. THE ANALYSIS METHOD

The analysis method is divided primarily into two parts, formulation and solution. (a) In the formulation part the LW [1] integration method is used to generate at each time step a set of nonlinear algebraic equations. (b) Then, this set is solved by a controlled-step [4] Newton's method.

(a) Formulation:

Consider the LW [1] integration formula

$$x_{n+1} = x_n + h[(1-\mu)\dot{x}_{n+1} + \mu\dot{x}_n] \tag{1}$$

with which is associated the error term,

$$e(t) = -h^2 \int_o^1 (\theta-\mu)\,\ddot{x}(t+\theta h)\,d\theta \tag{2}$$

where h is the time step size, $x_n \triangleq x(t_n)$, and μ is a free parameter with two important properties: (i) Equation (1) is A-stable* in the sense of

---

[1] This work was in part supported by Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Grant AF-AFOSR-71-2103 and in part supported by IBM, Department 688, Port Ewen, New York.

* An integration formula is A-stable if all numerical solutions, x_n, of the difference equation obtained by applying this formula to the scalar equation $\dot{x} = -\lambda x$ satisfy $|x_n| \to 0$ as $n \to \infty$ for any fixed h and any complex λ with $\mathrm{Re}\lambda > 0$.

Dahlquist [5] if and only if $\mu \in [0,\frac{1}{2}]$. (ii) Consider the scalar homogeneous equation $\dot{x} = -\lambda x$, $(\mathrm{Re}\lambda > 0)$, whose solution is $x(t_{n+1}) = e^{-q}x(t_n)$, $(q = \lambda h)$, then, the error incurred in simulating the amplification factor, e^{-q}, using (1) is

$$e(q) = ((1-\mu)/[1 + (1-\mu)q])-e^{-q} \tag{3}$$

If (3) is set equal to zero and solved for μ one gets

$$\mu(q) = \frac{1}{q} - \frac{1}{(e^{q}-1)} \tag{4}$$

LW showed that (a) $\mu(q)$ is a monotonically decreasing function of q for all real $q > 0$, and (b) for any given real λ there exists a μ equal to the $\mu(q)$ which lies in interval $\in [0,\frac{1}{2}]$ and thus is compatible with the stability condition. Conditions (i) and (ii) are used to preserve the numerical stability of formula (1), and give it flexibility in efficiently integrating a stiff* system of first order ordinary differential equations. Furthermore, bounds on the local truncation error of the LW method are given [6] as

$$|e(\mu)| \le h^2(\tfrac{1}{2}-\mu)|\ddot{x}(t + h)| + h^3M(4/3\,\mu^3 - \mu/2 + 1/6) \tag{5}$$

$$0 \le \mu \le \tfrac{1}{2}$$

$$M \equiv \max_{0 \le \theta \le 1} |\dddot{x}(t + \theta h)|$$

Special cases of (5) are

$$|e(o)| \le \frac{1}{2} h^2|\ddot{x}(t + h)| + \frac{1}{6} h^3M \qquad \text{(backward Euler)}$$

$$|e(\tfrac{1}{2})| \le \frac{1}{12} h^3M \qquad \text{(trapezoidal rule).}$$

With this in the background, we proceed with the formulation. Let $q^c_{n+1} = c^c_{n+1}\, e^c_{n+1}$ be the charge on a capacitor, c. The values of q^c_{n+1}, c^c_{n+1}, e^c_{n+1} are approximated at the (n+1)st time step. Using (1) and collecting terms we have

$$i_{n+1} = \left[\frac{1}{h(1-\mu)}\, c_{n+1} \right] e_{n+1} - \left[\frac{1}{h(1-\mu)}\, c_n e_n + \frac{\mu}{1-\mu}\, i_n \right]$$

$$= g^c_{n+1}\, e^c_{n+1} - I^c_n \tag{6}$$

where, the superscripts indicate the branch type, and

$$g^c_{n+1} \equiv \frac{1}{h(1-\mu)}\, c_{n+1} \tag{6-a}$$

$$I^c_n \equiv \frac{c_n}{h(1-\mu)}\, e_n + \frac{\mu}{1-\mu}\, i_n \tag{6-b}$$

The dual situation for the inductive branch, ℓ, is

$$e^{\ell}_{n+1} = r^{\ell}_{n+1}\, i^{\ell}_{n+1} - E^{\ell}_n \tag{7}$$

$$r^{\ell}_{n+1} \equiv \frac{1}{h(1-\mu)}\, \ell_{n+1} \tag{7-a}$$

---

*Let $\dot{x} = f(t,x)$, with x being its particular integral, be a system of first order ordinary differential equation, then this system is stiff if the eigenvalues of the Jacobian, $\partial f/\partial X$, are widely separated. [1]

$$E_n^{\ell} \equiv \frac{\ell_n}{h(1-\mu)} i_n + \frac{\mu}{(1-\mu)} e_n \tag{7-b}$$

Eqs. (6) and (7) give rise to the capacitor and inductor models of Figs. (1) and (2), respectively. These two branch models are special cases of the Kron's [7] multi-element branch of Fig. (3). From here on, all the quantities involved are assumed to have the (n+1)st subscript, unless they are explicitly defined otherwise. Let X be the vector of node voltages, E and I contain independent voltage and current sources and sources introduced due to modeling of reactive elements. Let G be the square matrix defined by Ohm's law, i.e., J = GV, where J and V are current and voltage vectors of the passive parts of all branches (see Kron's branch). To begin with, Kirchoff's current law is written as $A^t i(X) = 0$, where A is the incidence matrix, and i(X) is the vector of branch currents. Let $f(X) = A^t i(X)$ be the error vector measuring the discrepancy in satisfying the Kirchoff's current law. Referring to the Kron's branch,

$$\begin{aligned} f(X) &= A^t(J(X) - I) \\ &= A^t(J_1(X) + J_2(X) - I), \end{aligned} \tag{8}$$

where $J(x) \equiv J_1(X) + J_2(X)$ and $J_1(X)$ and $J_2(X)$ are associated with linear and nonlinear branches respectively. Eq. (8) can be written more explicitly if we apply Ohm's law to the linear branches, and note from the Kron's branch that

$$V = e + E = AX + E \tag{9}$$

where e = AX by nodal transformation, thus

$$f(X) = A^t[G_\ell(AX + E) + J_2(X) - I]. \tag{10}$$

Then, the Jacobian of f(X), which is needed to solve (10) for X can be computed as follows

$$\begin{aligned} \frac{\partial f(X)}{\partial X} &= A^t\, G_\ell A + A^t\, \frac{\partial J_2(X)}{\partial V} \cdot \frac{\partial V}{\partial X} \\ &= A^t\, G_\ell A + A^t\, \frac{\partial J_2(X)}{\partial V} A \end{aligned} \tag{11}$$

where, in (11) $A^t\, G_\ell A$ is the linear part of the Jacobian. A different derivation of (11) can be found in [8] for the nonlinear dc case.

(b) Solution - the Controlled Step-Newton's Method [4]:

Let $f = (f_1, f_2, \ldots, f_n)^t$ be a function mapping, mapping $X = (x_1, x_2, \ldots, x_n) \in S \to f(X) = (f_1(X), f_2(X), \ldots, f_n(X))^t \in T$. S and T are subsets of the Euclidean n-space, R^n. Consider the following modification of Newton's method for finding $X^* = (x_1^*, x_2^*, \ldots, x_n^*) \in S \ni \|f(X^*)\| \triangleq 0$. We pick $X_o \in S$ and recursively compute X_{i+1} from X_i as follows.

$$X_{i+1} = X_i + \lambda_i\, \Delta X_i \tag{12}$$

$$\frac{\partial f(X_i)}{\partial X_i}\, \Delta X_i = -f(X_i) \tag{13}$$

About this method we prove the following,

(i) $$\frac{d}{d\lambda_i} \|f(X_i + \lambda_i\, \Delta X_i)\|^2 \Bigg|_{\lambda_i = 0} = -2\|f(X_i)\|^2$$

(ii) If $f(X_i) = AX_i - b$, (A,b constant), then

$$\|f(X_i + \lambda_i \Delta X_i)\|^2 = (\lambda_i - 1)^2 \|f(X_i\|^2$$

To prove (i) and (ii), we expand $\|f(X_i + \lambda_i \Delta X_i)\|$ by a Taylor's series around $\lambda_i = 0$ [9].

$$\|f(X_i + \lambda_i \Delta X_i)\|^2 = \|f(X_i)\|^2 + \lambda_i (\nabla \|f(X_i)\|^2 \Delta X_i + \frac{1}{2} \lambda_i^2 \Delta X_i^t [\nabla \cdot (\nabla \|f(X_i)\|^2)^t] \Delta X_i + \ldots \tag{14}$$

Noting that

$$\nabla \|f(X_i)\|^2 = 2\left(\frac{\partial f(X_i)}{\partial X_i} \right)^t f(X_i) \tag{15}$$

and using (13) and taking the derivativw of (14) with respect to λ_i and evaluating it at $\lambda_i = 0$, (i) follows immediately. Using (13) and (15) we note that

$$\frac{1}{2} \Delta X_i^t [\nabla \cdot (\nabla \|f(X_i)\|^2)^t] \Delta X_i = 1 + (\Delta X_i^t H \Delta X_i), \tag{16}$$

where the kj-th element of H is $f^t(X_i) \frac{\partial^2}{\partial x_k x_j} f(X_i)$. Thus, if $f(X_i) = AX_i - b$, (16) reduces to 1. All higher terms in (14) reduce to zero (provided all partials of $f(X_i)$ exist). And in view of (i), (ii) follows. A more thorough proof of (i) and (ii) was first given by Zein and Hakimi [10]. A strategy for finding λ^* for error reduction is given by Branin [4].

Now, we are prepared to give an algorithm for solving the nonlinear transient problem. If n is the time step number, and i is the Newton's iteration number, then the algorithm consists of three main steps: Step 1 involves the outermost loop for advancing the time step. Step 2 concerns the intermediate loop for advancing the Newton's iteration number. And Step 3 has to do with an innermost loop for the application of a search strategy for λ^*. The algorithm in details is as follows.

1. Set $n = 0$.
2. Set $i = 0$.
3. Pick an initial guess X_i as follows: For $n = 0$, solve the dc initial condition problem by setting $\mu = 0$, $h \gg 1$. (The effect of this is that the capacitors and inductors are represented by small conductors and small resistors respectively, according to Eqs. (6) and (7).) Then, set $X_o = 0$. For $n > 0$, use a reasonable h, μ is updated using (4) to simulate rapidly changing exponentials. And $X_o = X_n$.
4. Evaluate V using (9), update nonlinearities in G and $J_2(X)$. Compute $f(X_i)$ and its Euclidean norm $\|f(X_i)\|^2$.
5. If $\|f(X_i)\|^2 < 10^{-9}$ go to 9, otherwise, go to 6.
6. Compute Jacobian using (11)*, solve for ΔX_i using (13)
7. Use a search strategy in the controlled-step Newton's method to find λ^*. If the search strategy does not result in error reduction, go to 2, else, go to Step 8.
8. Increase i by 1 and go to 4.
9. Print output at (n+1)st time step. Check if $t \geq t_f$, the final integration time, if so exist, otherwise, increase n by 1 and to to 2.

3. SENSITIVITY COMPUTATIONS - THE SENSITIVITY MATRIX

The construction of the gradient vector using parameter sensitivities is the most important step in the design and optimization of circuit responses. Methods, such as the Fletcher Powel [11], and the steepest descent [12], rely heavily in their success on the accuracy of the gradient vector. Because, the gradient vector in an optimization method plays the same important role as the correction vector of equation (13) in the controlled-step Newton's method. Therefore, we give below a simple easy-to-impliment and very efficient method for calculating parameter sensitivities. Our method, although restricted here to the node method of analysis, is applicable in an analogous fashion to the mesh, cutset and hybrid methods of formulation. Also, generalization to multi-terminal devices will be covered in a future paper.

---

*If h and μ are kept constant, the linear part of the Jacobian is computed only twice, once for initial condition solution, the other for the rest of the TR solution.

To compute the sensitivity matrix, we begin by taking differentials of both sides of (10) and in view of (9) and the fact that df(x) = 0 at convergence, then, collecting terms, we get,

$$[A^t G_\ell A + A^t \frac{\partial J_2(X_i)}{\partial V} A]\, dX_i = A^t[dI - G_\ell\, dE - dG_\ell V] \qquad (17)$$

or

$$B\, dX = A^t[dI - G_\ell dE - dG_\ell V]$$

$$B \equiv A^t G_\ell A + A^t \frac{\partial J_2(X)}{\partial V} A \quad \text{is the same Jacobian defined in (11),}$$

evaluated at convergence. Now, consider finding $\frac{\partial X}{\partial p_j}$. We set $dp_k = 0$ for all $k \neq j$ in (17) and pull dp_j* as a scalar from the right-hand side of (17) to get $B \frac{\partial X}{\partial p_j} = A^t a_j \epsilon_j$, where,

$$\epsilon_j = [0\ 0 \ldots \underset{\substack{\downarrow \\ \text{j-th entry}}}{1} \ldots 0]^t$$

or

$$\frac{\partial X}{\partial P_j} = B^{-1} a_j A^t_{j\cdot} = B^{-1} \underline{b}_j \qquad (18)$$

Where, in (18), a_j is a scalar that can be uniquely specified, $A_{j\cdot}$ is the j-th row of the incidence matrix, A, and $\underline{b}_j \equiv a_j A^t_{j\cdot}$. Compounding all the equations (18) for all j we get the sensitivity matrix,

$$\frac{\partial X}{\partial P} = B^{-1}[\underline{b}_1\ \underline{b}_2 \ldots \underline{b}_k] \qquad (19)$$

where, $P = (p_1, p_2, \ldots, p_k)^t$, and

$X = (x_1, x_2, \ldots, x_m)^t$ are a circuit parameter vector and the node voltage vector respectively. If sparse matrix methods are used it is more convenient to write (19) in the form

$$LU \frac{\partial X}{\partial P} = [\underline{b}_1\ \underline{b}_2 \ldots \underline{b}_k] \qquad (20)$$

where, B = LU, L and U are the lower and upper triangular matrices resulting from the decomposition of B [13]. The scalar, a_j, for all j is derivable from (17). We give here few examples. In the dc, frequency (AC) and time (TR) domains, if p_i is the parameter in the i-th branch, which is either I_i, E_i,

---

*We are assuming the branch, j, containing P_j, is not involved in coupling with other branches. A more thorough treatment of the sensitivity subject that includes coupling will be presented in a future work. However, see example 1.

g_{ii} (i.e. the i-th diagonal entry in G), $r_{ii} = \frac{1}{g_{ii}}$, then a_i is 1, $-g_{ii}$, $-v_i$, $\frac{1}{r_{ii}^2} v_i$ respectively. (v_i is the i-th entry of V (see (9)). In the AC case, Y = G in (10), and, if $p_i = c_i, \ell_i$, then $a_i = -sv_i$, $\frac{1}{s\ell^2} v_i$, respectively. In the TR domain, this depends on the integration method used. The following are important consequences of (19), (20)

1. In an optimization process involving the computation of the gradient only Km^2 multications are needed to compute the sensitivity matrix using (20) or (19). This number is substantially reduced if L and U are kept sparse in view of (20).
2. Stiffness and step-size control in the transient problem present no difficulties in computing the sensitivity matrix.
3. Optimization of nonlinear parameters is feasible, since these behave as linear parameters at convergence.
4. Eqs. (19) and (20), give exact sensitivities in the AC and dc domains. The accuracy in the TR domain is tied in with the accuracy of the numerical integration method used.

Two examples are given below, the first is applicable to the dc; and, in our formulation, at a time step in the TR domain, the second is applicable to the AC domain. In both examples $\frac{\partial X}{\partial P}$ is exact. Equation (19) is used.

Example 1 dc/TR (See Fig. (4))

$$A = \begin{bmatrix} -1 & 0 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\Delta} \left[\begin{array}{c|c} (G_2 + G_4) & G_2 \\ \hline (G_2 - \beta G_3) & G_1 + G_2 + G_3 - \beta G_3 \end{array} \right] \begin{bmatrix} G_1 E \\ 0 \end{bmatrix}$$

$$\text{or } X = B^{-1} Z$$

$$\Delta = G_1 G_2 + G_2 G_3 + G_1 G_4 + G_2 G_4 + G_3 G_4 (1-\beta)$$

Take $P = (E \; G_1 \; G_2 \; \beta \; G_3 \; G_4)^t$, then

$$\frac{\partial X}{\partial P} = B^{-1} \left[\begin{array}{c|c|c|c|c|c} G_1 & v_1 & v_2 & G_3 v_3 & (\beta-1) v_3 & 0 \\ \hline 0 & 0 & -v_2 & -G_3 v_3 & -\beta v_3 & v_4 \end{array} \right]$$

Example 2 AC (See Fig. (5))

$$A = \begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} (sC + \frac{1}{sL}) & sC \\ sC & G_1 + sC \end{bmatrix} \begin{bmatrix} G_1 E \\ 0 \end{bmatrix}$$

$$\text{or } X = B^{-1} Z$$

$$\Delta = G_1 sC + \frac{G_1}{sL} + \frac{C}{L}$$

Take $P = (E \; R_1 \; C \; L)^t$, then

---

[*] In the TR and AC domains, the number is per time step and frequency respectively.

$$\frac{\partial X}{\partial P} = B^{-1}\left[\begin{array}{c|c|c|c} G_1 & -\frac{v_1}{R^2} & -sv_2 & 0 \\ \hline 0 & 0 & sv_2 & \frac{1}{sL^2}\,v_3 \end{array}\right]$$

4. APPLICATIONS

A fairly gereral purpose FORTRAN computer program has been written. The program uses the algorithm given in Section 2. It also uses a controlled-step gradient method to perform optimization. Eq. (20) is made use of in the construction of the gradient. The whole program consists of 27 routines and functions, and is stored on disk as a permanent file. Only pertinent data is needed to analyze and/or optimize an arbitrary circuit. The program size is 9000 octal words (excluding system routines). A "COMMON" data area of 12,000 words is used for data. This allows a maximum problem size of 200 branches. Most of the 27 routines receive their arguments through "COMMON". The program uses sparse matrix methods. This means (a) only the nonzero entries of matrices are stored and processed, and (b) computation of the correction vector, using (13) is done by Gauss elimination, the LU decomposition [13], and sparse matrix technique with optimal ordering. The first optimal ordering algorithm of Tinney-Walker [14] is used. The storage format for matrices is basically that which is used in ECAP II [15], with a slight modification. The program architecture admits the addition of a language and dynamic storage allocation (DSA) system with minor modifications.

To test the efficiency of our analysis/optimization program we give three sample problems. The analysis-optimization data for the three examples given below are self-explanatory.

Example 1 Linear TR circuit (See Fig. (6))

(a) Analysis: h = 1 ns. Number of integration steps = 75, CPU time (CDC 6400) = 5.66 sec. Only two LU decompositions of the linear Jacobian were needed for analysis, the first one is for the dc initial condition solution, the second, for the rest of the transient solution.

(b) Optimization: $P = (R_2\ R_4\ R_6\ R_7\ R_{13})^t, P^* = (22.004k\ 1.48634k\ 2.91187k\ 5.958486k\ 3.275822k)^t$ = optimal parameter vector. Initial error (sum of squares) = 12.41128. Final error = .2875737. Error reduced to its minimum in one iteration, requiring 15 sec. of CPU time.

Example 2 Nonlinear biasing circuit (See Fig. (7))

Case #	P^o	e^o	E^o	P^*	e^*	E^*
1	$(R_1\ R_2\ R_4)^t =$ $(13\ 4\ 1)^t$	1.097	.182	$(12.9\ 4.38\ .32)^t$	1.48	$5.7x10^{-4}$
2	$(R_1\ E_1\ R_2\ R_4)^t =$ $(13\ 1\ 4\ 1)^t$	"	"	$(12.9\ 1.7\ 4.4\ .5)^t$	1.47	$9.9x10^{-4}$
3	$(R_1\ R_2\ R_4\ E_2)^t =$ $(13\ 4\ 1\ -2.5)^t$	"	"	$(12.91\ 4.37\ .3\ -2.83)^t$	"	$1x10^{-4}$
4	$(R_1\ E_1\ R_2\ R_4\ E_2)^t =$ $(13\ 1\ 4\ 1\ -2.5)^t$	"	"	$(12.91\ 1.64\ 4.35\ .54$ $-2.77)^t$	"	$1x10^{-4}$
5	$(G_1\ G_2\ G_4)^t =$ $(.077\ .25\ 1)^t$	"	"	$(.14\ .22\ 1)^t$	1.51	$4.1x.1^{-5}$

If $p_i \notin P^o$, its value is in the set $\{R_1,E_1,R_2,R_3,R_4,R_5,E_2\}$ = $\{13,1,4,6\text{x}10^{-5},1,6\text{x}10^{-5},-2.5\}$. P^o,P^* are initial and optimal parameter vectors. E^o,E^* are initial and minimum sum of squares error respectively. e^o,e^*,e^d = 1.5v. are initial, optimal and desired responses respectively. Convergence criterion = $1\text{x}10^{-3}$. (Met in all cases.) Time for case 5, 3.75 sec.

Example 3 Nonlinear TR problem - 'AND' gate (See Fig. (8))

(a) Analysis: No. of time steps = 100. $h = .25 \times 10^{-6}$. No. of controlled-step Newton's iterations for dc-initial-condition solution (with $X_o = 0$) = 5. Total Newton's iterations = 55. Time (CDC 6400) = 5.106 sec.

(b) Optimization: $P^o = (R_1\ R_2\ R_3\ E_2)^t = (1\ 1\ 10\ -2)^t$. $P^* = (1.030668\ .3316916\ 9.995671\ -4.655098)^t$. Initial and minimum sum of squares errors are 4.815020, .8037021 respectively. CPU time (CDC 6400) = 38.566 sec. (All resistors have been divided by 100. The capacitor and the nonlinear diode currents have been multiplied by 100.)

REFERENCES

[1] W. Liniger and R. A. Willoughby, "Efficient numerical integration of stiff systems of ordinary differential equations", IBM Res. Rep., RC 1970, December 20, 1967.

[2] D. A. Calahan, "Numerical consideration in the implimentation of a non-linear transient circuit analysis program", IEEE Trans. Circuit Theory, Vol. CT-10, pp. 66-72, January 1971.

[3] G. W. Gear, "Numerical integration of stiff ordinary differential equations", University of Illinois, Urbana, Rep. 221, January 1967.

[4] F. H. Branin,Jr., "A controlled-step Newton's method for solving nonlinear equations", Proc. ACM-SIAM-IEEE Conference on Mathematical and Computational Aids to Design, Anaheim, California, October 1969.

[5] G. G. Dahlquist, "A special stability criterion for linear multistep methods", BIT 3, 22-43 (1963).

[6] D. A. R. Zein, "Comments on "On the time-domain analysis of time invariant networks with large time constant spread"", (to appear in IEEE Circuit Theory, Vol. CT-19, March 1972.)

[7] G. Kron, Tensor Analysis of Networks, New York: Wiley, 1939.

[8] F. H. Branin, Jr. and H. H. Wang, "A fast reliable iteration method for dc analysis of nonlinear networks", Proc. IEEE, pp. 1819-1826, November 1967.

[9] T. M. Apostol, Mathematical Analysis, Addison-Wesley, Reading, Mass., 1957.

[10] D. A. R. Zein, "A research proposal for the degree of doctor of philosophy", Appendix III, Dept. of Electrical Engineering, Northwestern University, January 1971.

[11] R. Fletcher and M. J. D. Powel, "A rapidly convergent descent method for minimization", Computer J., 6 (1963), 163.

[12] J. Kowalik and M. R. Osborne, "Methods for unconstrained optimization problems", America Elsvier Publishing Co., New York, N. Y., 1969.

[13] G. E. Forsythe and C. B. Moler, "Computer solution of linear algebraic systems", Englewood Cliffs, N. J.: Prentice-Hall, 1967.

[14] W. F. Tinney and J. W. Walker, "Direct solutions of sparse network equations by optimally ordered triangular factorization", Proc. IEEE, Vol. 55, pp. 1801-1807, November 1967.

[15] F. H. Branin, G. R. Hogsett, R. L. Lunde, L. E. Kugel, ECAP II - An electronic circuit analysis program", IEEE Spectrum, June 1971.

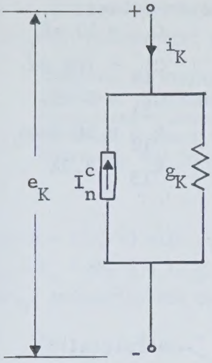


Fig. (1)
Capacitor model

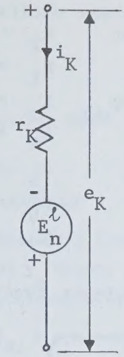

+
i_K
r_K
-
E_n^ℓ
+
e_K

Fig. (2)
inductor model

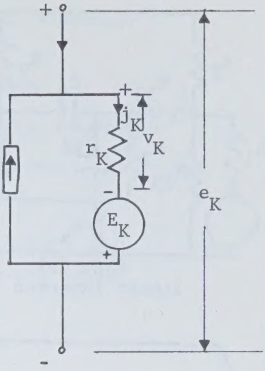

+
+
j_K
v_K
r_K
-
E_K
+
e_K
-

Fig. (3)
Kron's multi-element branch

Models at (n+1)st time step

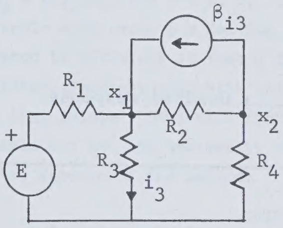

β_{i3}
R_1
x_1
R_2
x_2
+
E
R_3
i_3
R_4

Fig. (4)
dc/TR sensitivities

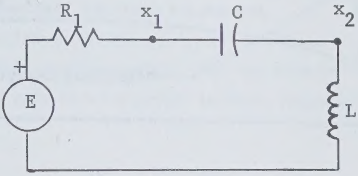
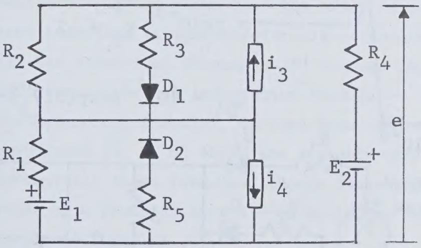


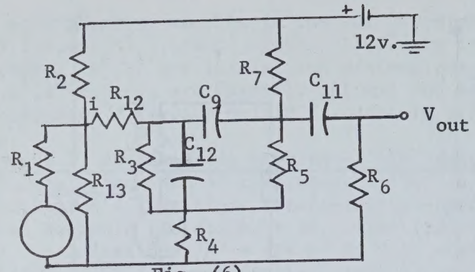
Fig. (5)
AC sensitivities


R_2
R_3
D_1
i_3
R_4
e
R_1
D_2
+
E_1
R_5
i_4
E_2
+

$$i_{Dj} = 10^{-9}(\exp(40v_j)-1);\ j = 1,2$$

$$i_3 = .98i_2,\ i_4 = .5i_1$$

Fig. (7)
nonlinear biasing circuit

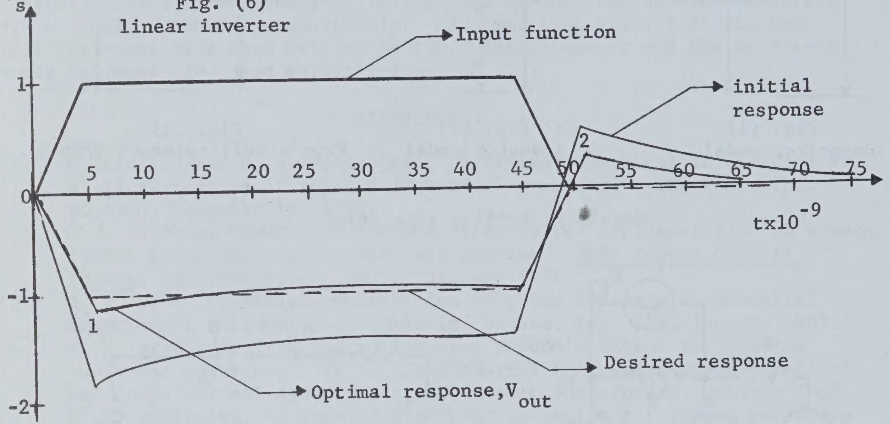


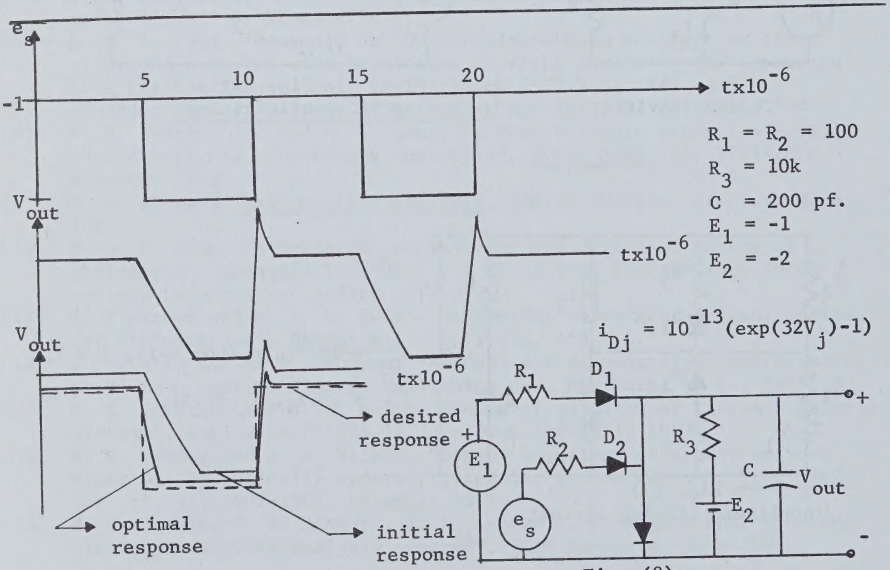
12v.
R2
R7
i
R12
C9
C11
Vout
R1
R3
C12
R5
R6
R13
R4

Fig. (6)
linear inverter

Initial parameter values:

R_1 = 1.1k
R_2 = 22k
R_3 = 150 ohms.
R_4 = 1.01k
R_5 = 30k
R_6 = 3k
R_7 = 6k
C_9 = 10 pf.
C_{10} = 100 pf.
C_{11} = 5 uf.
R_{12} = 50 ohms.
R_{13} = 3.3k





es
5
10
15
20
tx10^-6
-1
Vout
tx10^-6
Vout
tx10^-6
desired response
optimal response
initial response
R1
D1
+
R2
D2
R3
E1
es
C
Vout
E2
-

Fig. (8)
Nonlinear 'AND' gate

$R_1 = R_2 = 100$
$R_3 = 10k$
$C = 200$ pf.
$E_1 = -1$
$E_2 = -2$

$i_{Dj} = 10^{-13}(\exp(32V_j)-1)$

HIGH ORDER A-STABLE COMPOSITE MULTISTEP METHODS FOR NUMERICAL INTEGRATION OF STIFF DIFFERENTIAL EQUATIONS

THEODORE A. BICKART
DONALD A. BURGESS
Electrical & Computer Engineering Dept.
Syracuse University
Syracuse, New York

HARRY M. SLOATE
General Electric Electronics Laboratory
Syracuse, New York

ABSTRACT

Let $\dot{y} = f(y, t)$ with $y(t_o) = y_o$ possess a solution $y(t)$ for $t \geq t_o$. Set $t_n = t_o + nh$, $n = 1, 2, \ldots$. Let y_n denote the approximate solution of $y(t_n)$ defined by the composite multistep method

$$\begin{aligned} y_{N\ell} &= y_{N\ell-\ell} + h[\beta_{1\,\ell-1}\,\dot{y}_{N\ell} + \cdots + \beta_{10}\,\dot{y}_{N\ell-\ell}] \\ y_{N\ell-1} &= y_{N\ell-\ell} + h[\beta_{2\,\ell-1}\,\dot{y}_{N\ell} + \cdots + \beta_{20}\,\dot{y}_{N\ell-\ell}] \\ &\;\;\vdots \\ y_{N\ell-\ell+1} &= y_{N\ell-\ell} + h[\beta_{\ell\,\ell-1}\,\dot{y}_{N\ell} + \cdots + \beta_{\ell 0}\,\dot{y}_{N\ell-\ell}] \end{aligned}$$

with $\dot{y}_n = f(y_n, t_n)$ and $N = 1, 2, \ldots$. It is conjectured that the method can be A-stable with order $p = \ell+1$ for all $\ell \geq 1$ and shown to be so for $\ell = 1, \ldots, 6$. The method is efficient in that ℓ future approximate solution values are established simultaneously with only one function evaluation for each of the ℓ future time points. Step-size control is easily implemented in that the approximate solution and its derivative at only one past point appears in each component multistep formula of the method.

INTRODUCTION

Many dynamical systems are characterized by stiff differential equations. For example, a linear network with a large ratio of maximum to minimum moduli of natural frequencies is stiff. To achieve efficient numerical integration of stiff differential equations, it is necessary to choose a numerical integration method with care. One property sought in selecting a method is A-stability.

Dahlquist [1] has shown that the order of a multistep formula — an efficient numerical integration formula — cannot exceed 2 if it is A-stable. Ehle [2] has shown that Butcher's [3] ℓ-stage implicit Runge-Kutta formula — an inefficient numerical integration formula — of order 2ℓ is A-stable. Explicit Rosenbrock formulas, derived from semi-implicit Runge-Kutta formulas, have been found [4, 5, 6] which are A-stable with order at least as great as 4; unfortunately these formulas, though reasonably efficient, require precise evaluation of a Jacobian matrix — a stringent condition not imposed by the aforementioned formulas.

In this paper composite multistep methods — methods as efficent as multistep formulas — will be presented which are A-stable of order 2 through 7.

COMPOSITE MULTISTEP METHODS

Consider the differential equation

$$\dot{y} = f(y, t) \tag{1}$$

with the initial condition

$$y(t_o) = y_o. \tag{2}$$

Suppose (1) with (2) possesses a unique solution, $y(t)$, for $t \geq t_o$. Now, set $t_n = t_o + nh$. Then, let y_n denote the approximation of $y(t_n)$ defined for $n = k, k+1, \ldots$ by the composite multistep method*

$$\Sigma_{j=-k}^{\ell-1} [\alpha_{ij} y_{\ell N+m+j} - h\beta_{ij}\dot{y}_{\ell N+m+j}] = 0, \qquad i = 1, \ldots, \ell, \tag{3}$$

with $\dot{y}_n = f(y_n, t_n)$, $N = K, K+1, \ldots$, and $m = k - \ell K$ where K is the integer part of $(k + \ell - 1)/\ell$. In addition, it is assumed throughout that (3) possesses a unique solution for $h \in (0, \hat{h})$ and some $\hat{h} > 0$. Note: Although (1) will often be a vector differential equation — to which the results to follow will apply— for notational simplicity, only the scalar case will be considered initially.

Next, consider the expression

$$L_i[y(t_{\ell N+m-k}), h] = \Sigma_{j=-k}^{\ell-1}[\alpha_{ij} y(t_{\ell N+m+j}) - h\beta_{ij}\dot{y}(t_{\ell N+m+j})]. \tag{4}$$

Under the assumption that $y(t)$ in (4) possesses a convergent Taylor series at $t_{\ell N+m-k}$, (4) can be expressed as

$$L_i[y(t_{\ell N+m-k}), h] = \Sigma_{j=0}^{\infty} \gamma_{ij} y^{(j)}(t_{\ell N+m-k})\, h^j , \tag{5}$$

where

$$\gamma_{ij} = (1/j!)\, \Sigma_{h=-k}^{\ell-1} (k+n)^{j-1}[(k + n)\alpha_{in} - j\beta_{in}]. \tag{6}$$

The basis now exists for

DEFINITION 1: If p_i is an integer [the largest integer] such that $\gamma_{ij} = 0$ for $j = 0, \ldots, p_i$, then the i-th multistep formula of (3) is said to be of order [of exact order] p_i.

and

DEFINITION 2: Let $p = \mathrm{MIN}_{1 \leq i \leq \ell}\{p_i\}$, where p_i is the order [exact order] of the i-th multistep formula of (3). Then, the composite multistep method (3) is said to be of order [exact order] p.

The relationship between the order of a composite multistep method and its discretization error is: If the composite multistep method is of order p, if $y_{\ell N+m+j} = y(t_{\ell N+m+j}) + O[h^{p+1}]$ for $j = -k, \ldots, -1$, and if $y(t)$ is $p + 1$ times continuously differentiable for $t \geq t_o$, then $y_{\ell N+m+j} = y(t_{\ell N+m+j}) + O[h^{p+1}]$ for $j = 0, \ldots, \ell - 1$.

The set of ℓ scalar difference equations in (3) can be expressed as a single vector difference equation. Thus,

$$\Sigma_{j=-k}^{0}[\underset{\sim}{A}_j \underset{\sim}{Y}_{N+j} - h\underset{\sim}{B}_j \dot{\underset{\sim}{Y}}_{N+j}] = \underset{\sim}{0}, \tag{7}$$

where

$$\underset{\sim}{A}_j = \begin{bmatrix} \alpha_{1,(j+1)\ell-1} & \cdots & \alpha_{1,j\ell} \\ \vdots & & \vdots \\ \alpha_{\ell,(j+1)\ell-1} & \cdots & \alpha_{\ell,j\ell} \end{bmatrix}, \qquad \underset{\sim}{B}_j = \begin{bmatrix} \beta_{1,(j+1)\ell-1} & \cdots & \beta_{1,j\ell} \\ \vdots & & \vdots \\ \beta_{\ell,(j+1)\ell-1} & \cdots & \beta_{\ell,j\ell} \end{bmatrix},$$

---

*This is a somewhat restricted class of the composite multistep methods described in [7].

$$\underset{\sim}{Y}_{N+j} = \begin{bmatrix} y_{(N+j+1)\ell+m-1} \\ \vdots \\ y_{(N+j)\ell+m} \end{bmatrix}, \qquad \text{and} \qquad \dot{\underset{\sim}{Y}}_{N+j} = \begin{bmatrix} \dot{y}_{(N+j+1)\ell+m-1} \\ \vdots \\ \dot{y}_{(N-j)\ell+m} \end{bmatrix}.$$

Note that $\alpha_{ij} = 0$ and $\beta_{ij} = 0$ whenever $j < -k$. Furthermore, the underlying assumption that a unique solution exists for $h \in [0, \hat{h})$ implies that $\underset{\sim}{A}_o$ is nonsingular.

Let $\{y_n : n \geq \ell K+m = k\}$ denote the sequence defined by (3) or, equivalently, by (7). Now, consider

DEFINITION 3: Let $f(y, t) = qy$, where $\mathcal{R}e\{q\} < 0$. If $y_n \to 0$ as $n \to \infty$ for fixed $h > 0$, then the composite multistep method (3) is said to be A-stable.

Next, let

$$\underset{\sim}{R}(\zeta, \lambda) = \Sigma_{j=-k}^{0} \, [\underset{\sim}{A}_j - \lambda \, \underset{\sim}{B}_j] \, \zeta^{K+j} \tag{8}$$

Then it can be shown [7] that the composite multistep method is A-stable if and only if the moduli of the zeros of DET $\{\underset{\sim}{R}(\zeta, \lambda)\}$ as a polynomial in ζ do not exceed one.

It has elsewhere [7] been conjectured that the order of an A-stable composite multistep method cannot exceed 2ℓ. On the same basis of support established for that conjecture, it is tendered that the order of an A-stable composite multistep method cannot exceed 2 RANK $\{\underset{\sim}{B}_o\} \leq 2\ell$. This latter conjecture conforms to Dahlquist's result that an explicit multistep formula — composite multistep method with $\ell = 1$ and RANK $\{\underset{\sim}{B}_o\} = 0$ — cannot be A-stable.

A-STABLE METHODS WITH k = 1

The composite multistep methods with k = 1 — only one past point — are of particular value because, for them, step-size control routines are relatively simple. It has been shown [7] that the above conjecture is valid in this instance. In particular, the order of a composite multistep method with k = 1 cannot exceed $\ell + 1$.

Thusfar, it is known that A-stable methods with k = 1 and $p = \ell + 1$ exist for $\ell = 1, \ldots, 6$. Furthermore, some evidence exists in support of the belief that such methods exist for arbitrary ℓ.* For $\ell = 1$, the A-stable composite multistep method with k = 1 and $p = \ell + 1$ is the familiar, and A-stable, trapezoidal formula. The evolution of the method for $\ell = 2$ is presented next.

Each of the two component multistep formulas of the method must be of order 3; therefore, (6) yields, for i = 1 and 2,

$$\begin{aligned}
0 &= \alpha_{i1} + \alpha_{i0} + \alpha_{i,-1} \\
0 &= 2\alpha_{i1} + \alpha_{i0} - \beta_{i1} - \beta_{i0} - \beta_{i,-1} \\
0 &= 4\alpha_{i1} + \alpha_{i0} - 4\beta_{i1} - 2\beta_{i0} \\
0 &= 8\alpha_{i1} + \alpha_{i0} - 12\beta_{i1} - 3\beta_{i0}
\end{aligned}$$

---

*It is hoped that this suspected result will be validated and reported in the near future.

For each of the two values of i there are four linearly independent equations in six unknowns. Therefore, a solution can be obtained for four of the six unknowns in terms of the other two, here chosen to be α_{i0} and α_{i1}. The solution for i = 1 combined with that for i = 2 can be expressed in matrix form as

$$\begin{bmatrix} \alpha_{1,-1} & \beta_{11} & \beta_{10} & \beta_{1,-1} \\ \alpha_{2,-1} & \beta_{21} & \beta_{20} & \beta_{2,-1} \end{bmatrix} = - \begin{bmatrix} \alpha_{11} & \alpha_{10} \\ \alpha_{21} & \alpha_{20} \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 & 8 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & -4 & -12 \\ 0 & -1 & -2 & -3 \\ 0 & -1 & 0 & 0 \end{bmatrix}^{-1}$$

$$= \underset{\sim}{A}_0 \left\{ \frac{1}{12} \begin{bmatrix} -12 & 4 & 16 & 4 \\ -12 & -1 & 8 & 5 \end{bmatrix} \right\}$$

Note that the method is of order $\ell + 1 = 3$ for every choice of $\underset{\sim}{A}_0$ — previously constrained only to be non-singular.

The matrices $\underset{\sim}{A}_j$ and $\underset{\sim}{B}_j$ follow directly from the matrix of coefficients above; in addition to the arbitrary, nonsingular $\underset{\sim}{A}_0$ they are

$$\underset{\sim}{A}_{-1} = \underset{\sim}{A}_0 \begin{bmatrix} -1 & 0 \\ -1 & 0 \end{bmatrix}, \quad \underset{\sim}{B}_0 = \underset{\sim}{A}_0 \begin{bmatrix} 1/3 & 4/3 \\ -1/12 & 2/3 \end{bmatrix}, \quad \text{and } \underset{\sim}{B}_{-1} = \underset{\sim}{A}_0 \begin{bmatrix} 1/3 & 0 \\ 5/12 & 0 \end{bmatrix}.$$

The particular case of (8) is

$$\underset{\sim}{R}(\zeta, \lambda) = \underset{\sim}{A}_0 \left\{ \begin{bmatrix} 1 - \lambda/3 & -4\lambda/3 \\ \lambda/12 & 1 - 2\lambda/3 \end{bmatrix} \zeta + \begin{bmatrix} -1 - \lambda/3 & 0 \\ -1 - 5\lambda/12 & 0 \end{bmatrix} \right\}.$$

It then follows that

$$\mathrm{DET}\{\underset{\sim}{R}(\zeta,\lambda)\} = [\det\{\underset{\sim}{A}_0\}][\{\lambda^2/3 - \lambda + 1\}\zeta^2 + \{-\lambda^2/3 - \lambda - 1\}\zeta] .$$

The modulus of the single non-zero zero,

$$\zeta_1 = \frac{\lambda^2 + 3\lambda + 3}{\lambda^2 - 3\lambda + 3} ,$$

is easily shown to be less than one whenever $\mathcal{Re}\{\lambda\} < 0$; therefore, the method is A-stable. Note that ζ_1 does not depend upon $\underset{\sim}{A}_0$.

The evolution of a composite multistep method with k = 1 and $p = \ell + 1$ where ℓ is arbitrary, is similar. As for $\ell = 2$, the order is $\ell + 1$ for every choice of $\underset{\sim}{A}_0$ and the single non-zero zero of $\underset{\sim}{R}(\zeta, \lambda)$ as a polynomial in ζ does not depend on $\underset{\sim}{A}_0$. Also, as for $\ell = 2$, the modulus of the single non-zero zero will be found to be less than one for $\mathcal{Re}\{\lambda\} < 0$ and for $\ell = 3, \ldots, 6$; thus, the method is A-stable in these cases. Furthermore, each of the methods for a fixed choice of ℓ can be derived from any other such method — in particular, that for $\underset{\sim}{A}_0 = \underset{\sim}{I}$ — just premultiply the latter's matrix of coefficients by a suitable non-singular $\ell \times \ell$ matrix. Because of these observations the methods for which $\underset{\sim}{A}_0 = \underset{\sim}{I}$ will be considered canonical.* Now, the canonical composite multistep methods with k = 1 and $p = \ell + 1$ for 1 = 1,...,6 — those which are known to be A-stable — have the coefficients given in Table 1.

---

*The suggestion that the methods should be called canonical when $\underset{\sim}{A}_0 = \underset{\sim}{I}$ was made by William B. Rubin (Electrical & Computer Engineering Dept., Syracuse University).

TRUNCATION ERROR

With reference to (4), $L_i[y(t_{\ell N+m-k}), h]$ is the amount by which the solution of the differential equation (1) fails to satisfy the i-th component formula of the composite multistep method (3). As such it is a measure of the error between the solution of (1) and that of the i-th component formula of (3). Since the solution of the i-th component formula of (3) equals the truncated series of (1) through terms in h^p, it is known as the truncation error.

Let $Y = \mathrm{MAX}_{t \in [t_{\ell N+m-k},\ t_{\ell N+m+\ell+1}]} |y^{(p+1)}(t)|$. Then it can be shown that

$$|L_i[y(t_{\ell N+m-k}), h]| \leq h^{p+1} M_i Y. \tag{9}$$

Note: One procedure by which M_i can be determined is given in [8, pp. 246-248].

Now, let the vector $\underline{L}[y(t_{\ell N+m-k}), h]$ having $L_i[y(t_{\ell N+m-k}),h]$ as its i-th element be known as the truncation error vector of the composite multistep method (3). One measure of the error between the solution of (1) and that of (3) is the norm of the truncation error vector defined by

$$\|\underline{L}[y(t_{\ell N+m-k}),h]\| = \mathrm{MAX}_{1 \leq i \leq \ell} |L_i[y(t_{\ell N+m-k}), h]|. \tag{10}$$

By (9), this norm can be bounded as

$$\|\underline{L}[y(t_{\ell N+m-k}), h]\| \leq h^{p+1} MY \tag{11}$$

where $M = \mathrm{MAX}_{1 \leq i \leq \ell}\, M_i$.

Since the focus of this paper is on composite multistep methods with k = 1 and $p = \ell + 1$, the values of M for those methods will be established. Let $\mu = (t - t_{\ell N-\ell})/h$ and $\binom{\mu}{j} = [\mu(\mu - 1) \ldots (\mu - j + 1)]/j!$. Then, the Newton forward difference formula [9, p. 222], including the error term, for $\dot{y}(t)$ with the forward differences at $t_{\ell N-\ell}$ is

$$\dot{y}(t) = \Sigma_{j=0}^{\ell} \binom{\mu}{j} \Delta^j \dot{y}(t_{\ell N-\ell}) + \epsilon(\tau), \tag{12}$$

where $\Delta^j \dot{y}(t_{\ell N-\ell}) = \Delta^{j-1}\dot{y}(t_{\ell N-\ell+1}) - \Delta^{j-1}\dot{y}(t_{\ell N-\ell})$ is the j-th forward difference of $\dot{y}(t)$ at $t_{\ell N-\ell}$ and

$$\epsilon(\tau) = h^{\ell+1} \binom{\mu}{\ell+1} \dot{y}^{(\ell+2)}(\tau) \tag{13}$$

is the error term, τ being a function of t and in the largest closed interval containing $t_{\ell N-\ell}$, $t_{\ell N}$, and t.

Now, integrate (12) from $t_{\ell N-\ell}$ to $t_{\ell N+1-i}$ for $i = 1, \ldots, \ell$; the result is

$$y(t_{\ell N+1-i}) - y(t_{\ell N-\ell}) = h\, \Sigma_{j=0}^{\ell}\, \xi_{ij} \Delta^j \dot{y}(t_{\ell N-\ell}) + T_i\ , \tag{14}$$

where $\xi_{ij} = \int_0^{\ell+1-i} \binom{\mu}{j}\, d\mu$ and

$$T_i = h^{\ell+2} \int_0^{\ell+1-i} \binom{\mu}{\ell+1} y^{(\ell+2)}(\tau)\, d\mu. \tag{15}$$

Note: Since τ is a function of t and t is a function of μ, τ in this latter integral is a function of μ. Next, if the forward differences are expressed in terms of $\dot{y}(t_{\ell N-\ell})$ through $\dot{y}(t_{\ell N})$, it can be easily shown that (14) is equivalent to

$$\hat{L}_i[y(t_{\ell N-\ell}),h] = T_i, \tag{16}$$

where $\hat{L}_i[y(t_{\ell N-\ell}), h]$ is the truncation error associated with the i-th component multistep formula in the canonical composite multistep method for $k = 1$ and $p = \ell + 1$. Now, as a consequence of the fact that any other method can be derived from the canonical method by premultiplying the latter's matrix of coefficients by the former's $\underset{\sim}{A}_o$, the former's truncation error vector is

$$\underset{\sim}{L}[y(t_{\ell N-\ell}), h] = \underset{\sim}{A}_o \, \hat{\underset{\sim}{L}}[y(t_{\ell N-\ell}), h] = \underset{\sim}{A}_o \underset{\sim}{T} \tag{17}$$

where $\underset{\sim}{T} = [T_1 \; . \; . \; . \; T_\ell]'$. Note: Since $\underset{\sim}{A}_o = \underset{\sim}{I}$ for the canonical method, (17) also holds for that method.

It follows from (17) that

$$\|\underset{\sim}{L}[y(t_{\ell N-\ell}), h]\| \leq \|\underset{\sim}{A}_o\| \, \|\underset{\sim}{T}\| , \tag{18}$$

where the matrix norm is that which is subordinate to the vector norm. Thus, $\|\underset{\sim}{A}_o\| = \mathrm{MAX}_{1 \leq i \leq \ell} \sum_{j=o}^{\ell-1} |a_{ij}|$. From (15) by the usual inequality relations

$$|T_i| \leq h^{\ell+2}\{\textstyle\int_0^{\ell+1-i} |\binom{\mu}{\ell+1}| d\mu\} \, Y. \tag{19}$$

Therefore,

$$\|\underset{\sim}{T}\| \leq h^{\ell+2} \; \{\mathrm{MAX}_{1 \leq i \leq \ell} \int_0^{\ell+1-i} |\binom{\mu}{\ell+1}| d\mu\} \, Y = h^{\ell+2} \; \{\int_0^{\ell} |\binom{\mu}{\ell+1}| d\mu\} \, Y \tag{20}$$

Combine (18) and (20) and compare the result with (11); it follows that a suitable value of M is

$$M = \|\underset{\sim}{A}_o\| \int_0^{\ell} |\binom{\mu}{\ell+1}| \, d\mu . \tag{21}$$

Values of M conforming to (21) for the canonical methods — $\underset{\sim}{A}_o = \underset{\sim}{I}$ and $\|\underset{\sim}{A}_o\| = 1$ are listed in Table 2 for $\ell = 1, \ldots, 6$.

Another, and usually smaller, value of M can be established. This time, take the norm of (17) only after taking the matrix product indicated; the result is

$$\|\underset{\sim}{L}[y(t_{\ell N-\ell}), h]\| = \mathrm{MAX}_{1 \leq i \leq \ell} \, |\textstyle\sum_{j=1}^{\ell} a_{i,\ell-j} T_j| \tag{22}$$

By (15) and the usual inequality relations

$$|\textstyle\sum_{j=1}^{\ell} a_{i,\ell-j} T_j| \leq h^{\ell+2} \; \{\sum_{j=1}^{\ell} |a_{i,\ell-j}| \int_0^{\ell+1-j} |\binom{\mu}{\ell+1}| d\mu\} \, Y. \tag{23}$$

Now, combine (23) with (22) and compare the result with (11); it follows that a suitable value of M is

$$\begin{aligned} M &= \mathrm{MAX}_{1 \leq i \leq \ell} \sum_{j=1}^{\ell} |a_{i,\ell-j}| \int_0^{\ell+1-j} |\binom{\mu}{\ell+1}| d\mu \\ &= \mathrm{MAX}_{1 \leq i \leq \ell} \sum_{k=o}^{\ell-1} \{\sum_{j=k}^{\ell-1} |a_{ij}|\} \int_k^{k+1} |\binom{\mu}{\ell+1}| \, d\mu . \end{aligned} \tag{24}$$

The second line can be derived from the first by setting $\int_0^{\ell+1-j} |\binom{\mu}{\ell+1}| d\mu = \sum_{k=o}^{\ell-j} \int_k^{k+1} |\binom{\mu}{\ell+1}| d\mu$ and then interchanging the order of summation. Note: Evaluation of the second line of (24) can be made easier by observing that $\int_k^{k+1} |\binom{\mu}{\ell+1}| d\mu = |\int_k^{k+1} \binom{\mu}{\ell+1} d\mu|$. Values of the integrals in (24) are given in Table 3. Note that (21) and (24) yield the same value for M when $\underset{\sim}{A}_o = \underset{\sim}{I}$.

EXAMPLES

The canonical composite multistep method with $k = 1$ and $p = \ell + 1$ is the basis for a numerical integration process programmed in APL. The program incorporates a modified Newton method to solve the implicit algebraic multistep formulas of the composite multistep method. The modification eliminates the need to compute the Jacobian associated with the composite multistep method for every iteration at each block of ℓ steps. Using the truncation error, the step size is adjusted to as large a value as possible at each block of ℓ steps so as to bound the local discretization error.

The following three differential equations — the first two of which are stiff — were chosen and then solved with the program to illustrate the performance of the method as the basis for a variable-step numerical integration process:

$$\text{I.} \quad \dot{y} = -1000y + t \qquad y(0) = 3$$

$$\text{II.} \quad \dot{\underset{\sim}{y}} = \begin{bmatrix} -2000 & 1000 \\ 1 & -1 \end{bmatrix} \underset{\sim}{y} + \begin{bmatrix} 1000 \\ 0 \end{bmatrix} \qquad \underset{\sim}{y}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\text{III.} \quad \dot{y} = (2t - y)/(t + 2y) \qquad y(1) = 0$$

The initial step size, the final step size, and the number of blocks of ℓ steps are given in Table 4 for $\ell = 1, \ldots, 6$. In each case the final time was 5 and the local discretization error — relative for solution values exceeding unity and absolute otherwise — did not exceed the preset value of 10^{-5}. The important observation to make is: The step size in cases I and II — stiff differential equations — increased significantly as the variation in the solution decreased.

The time required to effect the solution has not been given as the program is still in the process of being modified to decrease its execution time.* Furthermore, a variable-order feature is being incorporated in the program. The resulting program will be released and reported shortly. A FORTRAN version will be available in the near future. At that time comparative tests using other variable-step variable-order processes will be reported.

ACKNOWLEDGEMENT

The research reported herein was supported by the National Science Foundation under Grant GK-23010.

REFERENCES

[1] G.C. Dahlquist, A Special Stability Problem for Linear Multistep Methods, B.I.T., 3(1963), pp. 27-43.

[2] B.L. Ehle, High Order A-Stable Methods for the Numerical Solution of Systems of D.E.'s, B.I.T., 8(1968), pp. 276-278.

[3] J.C. Butcher, Implicit Runge-Kutta Processes, Math. Comp., 18(1964),pp.50-64.

[4] H.H. Rosenbrock, Some General Implicit Processes for the Numerical Solution of Differential Equations, Comp. J., 5(1963), pp. 329-330.

[5] D.A. Calahan, A Stable, Accurate Method of Numerical Integration for Nonlinear Systems, IEEE Proc., 56(1968), p. 744.

[6] R.H. Allen and C. Pottle, Stable Integration Methods for Electronic Circuit Analysis with Widely Separated Time Constants, Proc. Sixth Allerton Conf. on Circuit and System Theory, Univ. of Illinois, Urbana, Illinois, 1968, pp. 311-320.

---

* Also, no other methods have yet been programmed in APL to provide a basis for comparing solution times.

[7] H.M. Sloate and T.A. Bickart, A-Stable Composite Multistep Methods, 1971 SIAM National Meeting

[8] P. Henrici, Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons, New York, 1962.

[9] P. Henrici, Elements of Numerical Analysis, John Wiley and Sons, New York, 1964.

$$\begin{bmatrix} \alpha_{1\,-1} & \beta_{1\,\ell-1} & \cdots & \beta_{1\,-1} \\ \vdots & \vdots & & \\ \alpha_{\ell\,-1} & \beta_{\ell\,\ell-1} & \cdots & \beta_{\ell\,-1} \end{bmatrix}$$

$\ell = 1$

$$\frac{1}{2}\begin{bmatrix} -2 & 1 & 1 \end{bmatrix}$$

$\ell = 2$

$$\frac{1}{12}\begin{bmatrix} -12 & 4 & 16 & 4 \\ -12 & -1 & 8 & 5 \end{bmatrix}$$

$\ell = 3$

$$\frac{1}{24}\begin{bmatrix} -24 & 9 & 27 & 27 & 9 \\ -24 & 0 & 8 & 32 & 8 \\ -24 & 1 & -5 & 19 & 9 \end{bmatrix}$$

$\ell = 4$

$$\frac{1}{720}\begin{bmatrix} -720 & 224 & 1024 & 384 & 1024 & 224 \\ -720 & -27 & 378 & 648 & 918 & 243 \\ -720 & -8 & 32 & 192 & 992 & 232 \\ -720 & -19 & 106 & -264 & 646 & 251 \end{bmatrix}$$

$\ell = 5$

$$\frac{1}{1440}\begin{bmatrix} -1440 & 475 & 1875 & 1250 & 1250 & 1875 & 475 \\ -1440 & 0 & 448 & 2048 & 768 & 2048 & 448 \\ -1440 & 27 & -189 & 1026 & 1026 & 1971 & 459 \\ -1440 & 16 & -96 & 224 & 224 & 2064 & 448 \\ -1440 & 27 & -173 & 482 & -798 & 1427 & 475 \end{bmatrix}$$

$\ell = 6$

$$\frac{1}{60480}\begin{bmatrix} -60480 & 17712 & 93312 & 11664 & 117504 & 11664 & 93312 & 17712 \\ -60480 & -1375 & 28200 & 58125 & 80000 & 31875 & 87000 & 18575 \\ -60480 & -512 & 3072 & 11136 & 96256 & 24576 & 89088 & 18304 \\ -60480 & -783 & 5832 & -19683 & 58752 & 31347 & 87480 & 18495 \\ -60480 & -592 & 4224 & -12912 & 21248 & 528 & 90240 & 18224 \\ -60480 & -863 & 6312 & -20211 & 37504 & -46461 & 65112 & 19087 \end{bmatrix}$$

Table 1 Coefficient matrices: Canonical composite multistep methods with $k=1$ and $p=\ell+1$.

$\ell \rightarrow$	1	2	3	4	5	6
$M \rightarrow$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{49}{720}$	$\frac{19}{360}$	$\frac{2459}{60480}$	$\frac{71}{2240}$

Table 2 Error coefficient M evaluated as in (21).

$k \downarrow$ \ $\ell \rightarrow$	1	2	3	4	5	6
0	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{19}{720}$	$\frac{3}{160}$	$\frac{863}{60480}$	$\frac{275}{24192}$
1		$\frac{1}{24}$	$\frac{11}{720}$	$\frac{11}{1440}$	$\frac{271}{60480}$	$\frac{13}{4480}$
2			$\frac{19}{720}$	$\frac{11}{1440}$	$\frac{191}{60480}$	$\frac{191}{120960}$
3				$\frac{3}{160}$	$\frac{271}{60480}$	$\frac{191}{120960}$
4	$\int_k^{k+1} \left\| \binom{\mu}{\ell+1} \right\| d\mu,\ k=0,\dots,\ell-1$				$\frac{863}{60480}$	$\frac{13}{4480}$
5						$\frac{275}{24192}$

Table 3 Integrals in error coefficient M evaluated as in (24).

PROBLEM	NUMERICAL SOLUTION PROPERTIES	ℓ 1	2	3	4	5	6
I	INITIAL STEP SIZE	4.44×10^{-5}	9.93×10^{-5}	1.64×10^{-4}	2.31×10^{-4}	2.96×10^{-4}	3.56×10^{-4}
	FINAL STEP SIZE	6.70	2.36	1.89	1.35	1.83	0.636
	NUMBER OF BLOCKS OF ℓ STEPS	78	32	19	15	13	11
II	INITIAL STEP SIZE	2.80×10^{-5}	5.90×10^{-5}	9.42×10^{-5}	1.30×10^{-4}	1.64×10^{-4}	1.94×10^{-4}
	FINAL STEP SIZE	0.205	0.321	0.486	0.566	0.643	0.761
	NUMBER OF BLOCKS OF ℓ STEPS	98	37	21	16	13	12
III	INITIAL STEP SIZE	0.0102	0.0154	0.0385	0.0400	0.0338	0.0399
	FINAL STEP SIZE	0.343	0.415	0.348	0.414	0.319	0.289
	NUMBER OF BLOCKS OF ℓ STEPS	56	23	11	9	7	6

Table 4 Examples: Numerical solution properties

A DIGITAL SIMULATION METHOD FOR INTERCONNECTED CONTINUOUS SYSTEMS *

RICHARD A. DIPERNA
Philips Industries
ISA Division
Croydon, England.

The design of a complex system is usually accomplished by partitioning the system into subsystems with well-defined tasks and interfaces. Consider, therefore, the simulation of a system formed by the interconnection of a collection of subsystems S_i, $i=1,\ldots,N$, with input vectors v_i and output vectors y_i. Assume that each subsystem input is connected either to an output of another subsystem or to an external input. Then, letting $\underline{v}(t)=\text{col}[v_1(t),\ldots,v_N(t)]$ and $\underline{y}(t)=\text{col}[y_1(t),\ldots,y_N(t)]$, the interconnection can be represented as:

$$\underline{v}(t) = M\,\underline{y}(t) + P\,\underline{u}(t) \qquad (1)$$

where M and P are topological matrices and where $\underline{u}(t)$ is a vector of external inputs. The system to be simulated is defined by equation (1) together with the dynamics of the subsystems.

One method commonly used to simulate such a system is to form the state equations of the system and to integrate these numerically. This method is especially suitable for systems which can be represented as a block diagram composed of integrators and function blocks. In this case, the state equations can be formed automatically by sorting the interconnection equations to obtain all integrator inputs, which are the desired state variable derivatives, as functions of external inputs and integrator outputs (state variables).

For many systems, however, it is difficult or impossible to formulate the state equations. For example, this is so if one or more subsystems are described only by impulse responses. Also, for those systems which can be represented as a block diagram of integrators and function blocks, the system equations may be of the form:

$$\begin{aligned} \dot{x} &= f(x,u) \\ 0 &= g(x,u) \end{aligned} \qquad (2)$$

where x is the state vector and u is a vector of auxiliary variables. If the second equation cannot be solved explicitly for u in terms of x, then the state equations are not directly available.

---

* This work was supported in part by the National Aeronautics and Space Administration under Grant NGR-33-0008-090, by the National Science Foundation under Grant GK-2283 and by the Armstrong Foundation at Columbia University, N.Y.,N.Y.

For these systems, therefore, a different approach is proposed. A discrete model should be generated for each individual subsystem and these models then interconnected to obtain the overall simulation. The mathematical nature of the interconnection process is determined by the numerical techniques used to model the subsystems. Recently, much attention has been focused on implicit numerical techniques because of their advantage in efficiency for the simulation of stiff systems, i.e., systems with widely varied time constants [1-2]. If a subsystem is modelled using implicit techniques, its output at time t_n will usually depend on its input at time t_n[3]. In this case, the equation for the subsystem model may be written in the following form, with the notation simplified by omitting all arguments not relevant to the discussion:

$$y_{i,n} = r_i(v_{i,n}, t_n) \quad i=1,\dots,N$$

or in vector notation,

$$\underline{y}_n = \underline{r}(\underline{v}_n, t_n)$$

where $\underline{r}=\mathrm{col}[r_1,\dots,r_N]$. If we substitute this into equation (1), the result is:

$$\underline{v}_n = M\,\underline{r}(\underline{v}_n, t_n) + P\,\underline{u}_n \tag{3}$$

Since $\underline{u}_n$ is the known vector of external inputs, the equation (3) represents a system of implicit equations in $\underline{v}_n$ which are usually sparse. These equations can be solved by iteration at each time point to yield the simulation results. However, rather than solving the complete set of equations, it is usually more efficient to reduce the dimensionality of the problem by algebraic substitution, which can be done automatically by sorting of the interconnection equations [3].

The computational cost in solving equation (3) at each step is justified for three reasons: (a) for many systems the simulation may not be possible otherwise; (b) for the case of systems with widely varied time constants, the increased efficiency more than compensates for the extra cost; (c) subsystems of the type described by equation (2) can be treated directly [3].

REFERENCES

[1] G.G. Dahlquist, 'A Special Stability Problem for Linear Multistep Methods,' BIT 3 no.1, pp.27-43 (1963).

[2] C.W. Gear, 'Numerical Integration of Stiff Ordinary Differential Equations,' University of Illinois, Urbana, Rep.221, (January 1967).

[3] R.A. DiPerna, 'Computational Methods for Digital Simulation of Continuous Systems,' Columbia University Tech.Rpt.123, Department of Electrical Engineering, Columbia University, New York, N.Y. (June 1970).

RECENT ADVANCES IN DATA COMMUNICATION

H. O. Burton
Bell Telephone Laboratories
Holmdel, N. J.

ABSTRACT

Advances in the facilities on which data are transmitted have been and will be significant. The capability of the nationwide DDD network for transmitting data is improving. Plans for completely new networks have been announced by MCI and DATRAN. Western Union and the Bell System are also planning new facilities for digital data transmission.

A SURVEY OF DATA COMPRESSION

LEE D. DAVISSON
University of Southern California, Los Angeles, California

I. INTRODUCTION

The purpose of this paper is to present a review of some of the important activities presently going on in the area of data compression in the perspective of past achievements. Data compression, sometimes called bandwidth compression, source coding or data compaction, is any method which is used to attempt to reduce the channel requirements for the transmission of a given source while maintaining acceptable data quality. In Figure 1 the problem can be stated as one of finding the best encoder for the given source, generating output message denoted X, for transmission over the channel so that the output message Y is a "good" reproduction of X. The decoder is usually easily determined once the encoder is given and hence does not require discussion. Note that the channel is assumed noiseless. This is done for the purposes of this paper to restrict the subject to the already difficult source encoding problem only. Thus, whatever channel coding is required to attain reliable transmission is contained within that block labeled "noiseless channel". The reader should keep in mind that in subsequent material, any errors referred to are due to the encoding procedure and not channel noise.

In succeeding sections results are given on the optimal encoding of sources. Because of the survey nature, many imprecise statements are made, mathematically and otherwise, for the sake of conveying the essential ideas. The author apologizes in advance.

II. RATE DISTORTION THEORY

Rate distortion theory as developed by Shannon [1] provides a theoretical basis for the coding of sources. It proceeds from the knowledge of a source probabalistic description of the message output, X, together with an appropriate measure, d(X, Y), called a distortion measure, of the relative importance of errors in the reconstructed message, Y, at the receiver output. The value of the rate distortion function is the minimum (noiseless) transmission requirement, i.e. channel capacity, for the source at a given level of average distortion and is mathematically:

$$R(D) = \inf_{E[d(X,Y)] \leq D} I(X,Y)$$

where $E[\cdot]$ denotes mathematical expectation and I(X, Y) is the mutual information between source message, X, and decoder output Y. The minimization is over all possible encoders where any given encoder-decoder combination is described by conditional probabilities of the decoder output given the encoder input.

Rate distortion theory has three fundamental limitations:

1) The source probabilistic description may be unknown, or at least only partially known.

2) The user may not be able to formulate precisely a distortion measure $d(\cdot, \cdot)$.

3) The mutual information is a difficult function to minimize over. Therefore few solutions of practical interest are known and those that are may not include the given source and the user's distortion measure, if known.

In spite of the preceding limitations, rate distortion theory can provide bounds on performance and justification and guidance as to the encoder form. At any rate work continues on the theory. An excellent presentation of current and past investigations is in the book of Berger [3]. Future results may lessen the importance of the above comments.

Consider a source generating time-discrete outputs. For data compression to be worthwhile, the source must have memory, i.e. statistical dependence between successive outputs, because it is here that encoding gains of a useful magnitude can be made. Unfortunately rate distortion theoretical results are particularly limited for sources with memory. One of particular theoretical interest [2, 3] is the rather remarkable result that for small enough (but nonzero) distortions, for distortion measured on an output-by-output basis, the rate distortion function for certain classes of sources with memory is identical with a particular memoryless source. Because the rate distortion function for the memoryless source can be found with relative ease, the rate distortion function for the source with memory also results. Unfortunately the answers found have not lead to practical methods of achieving the bound, nor has it yet been possible to extend the answers above some cutoff distortion level.

The other result for sources with memory to be discussed herein is for the Gaussian source with a quadratic distortion measure. Although many sources may not be Gaussian, they may be approximately so, and although a quadratic distortion measure may not be appropriate, the results indicate encoder configurations which seem reasonable for other distortion measures. This is discussed in the next section.

III. TIME-DISCRETE GAUSSIAN SOURCE WITH FREQUENCY-WEIGHTED DISTORTION MEASURE

Measuring average distortion by the average squared frequency error is not the most general quadratic distortion measure but in many practical applications has the most physical meaning. Given a source spectral density, $S(\lambda)$, and an error spectral density, $E(\lambda)$, at the decoder output, the mutual information is given by one-half the integral of the log of the power signal-to-error spectral density ratio. Measuring the average distortion by

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} W(\lambda)E(\lambda)d\lambda$$

where $W(\lambda)$ is a chosen frequency weight, the rate distortion function is given by

$$R(D) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\log\frac{S(\lambda)}{E(\lambda)}\,d\lambda$$

where in minimizing the mutual information,

$$E(\lambda) = \min(S(\lambda), \alpha/W(\lambda))$$

where α is a constant chosen so that the distortion level equals D. This result was first given by Kolmogorov [4]. Figure 2 demonstrates the situation pictorially. All frequencies where $\alpha/W(\lambda) \geq S(\lambda)$ are not transmitted at all but are replaced by zero resulting in a distortion identically equal to the source spectral density. All other frequencies are transmitted with signal-to-error power spectral density ratio $S(\lambda)/E(\lambda) = S(\lambda)W(\lambda)/\alpha$. Methods of approximately realizing the bound are discussed in the next section.

IV. BLOCK TRANSFORM ENCODING

Consider for the moment encoding to minimize mean square error ($W(\lambda)$ = constant). If the data is to be encoded in blocks of length N, then the minimum rate requirements are found in using the eigenfunctions of the N × N covariance matrix, Φ_N, of the source as a set of basis functions. The N-block rate distortion function is

$$R_N(D) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\log[\lambda_i/\min(D_1, \lambda_i)]$$

where $\{\lambda_i\}$ are the eigenvalues of Φ_N and D_1 satisfies

$$D = \frac{1}{N}\sum_{i=1}^{N}\min(D_1, \lambda_i) .$$

This can also be put in the form

$$R_N(D) = \frac{1}{2N}\log[|\Phi_N|] - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\log[\min(D_1, \lambda_i)]$$

$$= \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{2}\log\frac{\sigma_i^2}{\sigma_\infty^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\log[\min(D_1, \lambda_i)]$$

σ_i^2 is the one-step ahead prediction error of the source based on i past values.

Ignoring the second term due to the distortion, a measure of the loss in finite block encoding is the first term in the last equation which has to do with the "predictability" of the source which an experimentor can sometimes most easily estimate. For a Gauss-Markov source with correlation coefficient ρ, the loss measure becomes

$$\frac{1}{2N}\log\frac{1}{1-\rho^2} .$$

For two dimensional (i.e. picture) data the same type of calculations can be made, e.g. line-by-line encoding vs. whole picture coding[5].

Although the eigenfunctions are the best set, they require extensive calculations to find. The Fourier transform coordinates on the other hand are equivalent in the limit as N → ∞ and, since they are frequency coordinates, provide a natural set of basis functions on which to place a frequency weighted distortion measure. For a Gauss-Markov source, the N-block loss can be shown to be

$$\leq \frac{1}{2} \log \; 1 + \frac{1-\rho^N}{N} \rho \frac{1+\rho^2}{1-\rho^2} \text{ bits .}$$

which goes to zero as $N \to \infty$.

A third set of coordinates of practical interest is the Walsh functions. These do not, in general, approach zero loss as $N \to \infty$, but are the most simple to use computationally. For a Gauss-Markov source with $\rho = .95$ and $N = 512$, the loss has been calculated [6] as .298 bits. For the Fourier transform for the same values, the loss is practically zero.

IV. TIME-DOMAIN METHODS

If the system quantizer in Figure 2 were to insert Gaussian noise, then the rate distortion function would be achieved by the given configuration in the sense that the capacity of the quantizer (looked at as a channel) would equal the rate distortion function [6-9]. The filters are chosen to achieve any desired frequency weighting function. However, because practical quantizers do not generate Gaussian noise, one of the difficult long standing analytical questions is the effect of such a quantizer in the feedback loop on system performance [10].

Figures 4 and 6 present suboptimum configurations based on Figure 3. Figure 4 results by removing the pre and post filtering operations and is a form of predictive DPCM. If the quantizer is looked at as adding Gaussian noise, for a Gauss-Markov source with squared error distortion measure, the encoding loss can be shown to be

$$\leq \frac{1}{2} \log \frac{1}{1 - D}$$

which is negligible for small distortions. The configuration has the additional advantage of absolute as well as average-squared control over the error if a practical quantizer is used. In addition, the minimum mean square predictor can be made adaptive [10] reducing the need for source probabilistic information.

In Figure 6 a noiseless coder follows a quantizer which inserts the maximum allowable distortion. Figure 5 shows quantizing loss upper bound for a uniform step size quantizer as measured by the first order conditional entropy of the quantizer output as a function of the rate distortion function for the resulting distortion level when a two dimensional conditional mean reconstructor is used. Several values of the correlation coefficient are given and it is seen that the loss is not too severe for rates greater than one bit relative to the apparent inherent loss in uniform quantizing measured by the $\rho = 0$ curve. The value of around .25 bit seems to be the best that practical quantizers can achieve [11].

Quantized data can be coded by the methods of the next section.

V. UNIVERSAL CODING

For amplitude-discrete, time discrete data, combinatorial methods of encoding without error have been applied [12-14], sometimes called universal encoding. These methods enable one to transmit at rates approaching the source entropy asymptotically as the block size N goes to infinity without knowledge of the source output probabilities except that they are of some class (e.g. Mth order Markov). Essentially the idea is this. The source probability function for output X can be written as

$$P_\theta(X) = g_\theta(\theta_s(X))h(X)$$

where θ is an unknown parameter vector and $\theta_s(X)$ is a sufficient statistic for θ. Transmit $\theta_s(X)$ using a fixed length message requiring something proportional to log N bits. Optimally encode the rest using the known function h(X) requiring an asymptotic average of 2^{HN} bits where H is the source entropy per output symbol. Then the coder requires an average of H plus a term like log N/N bits which goes to zero with the block size N.

The idea can most easily be illustrated by simple example. Consider a source generating independent binary outputs with Prob(one) = p, Prob(zero) = 1-p. Send the following two part message:

(number of ones)(location of the ones in the block)

The first part requires log N fixed length messages. All location sequences of the ones given their number are equally probable so for M ones, log $\binom{N}{M}$ messages are required. It can be shown [12] that the number of bits per output,

$$\frac{1}{N}\left[\log N + \log\binom{N}{M}\right]_{N\to\infty} \to -p \log p + (1-p)\log(1-p) .$$

REFERENCES

1. C. E. Shannon, "Coding Theorems for a Discrete Source with a Fidelity Criterion," Information and Decision Processes, McGraw-Hill, 1960.

2. R. M. Gray, "Information Rates of Stationary Ergodic Finite Alphabet Sources," IEEE Transactions on Information Theory, Sept. 1971.

3. T. Berger, Rate Distortion Theory - A Mathematical Basis for Data Compression, Prentice-Hall, 1971.

4. A. N. Kolmogorov, "On the Shannon Theory of Information Transmission in the Case of Continuous Signals," IRE Trans. on Information Theory, Dec. 1956.

5. D. J. Sakrison and V. R. Algazi, "Comparison of Line-by-Line and Two-Dimensional Encoding of Random Images," IEEE Trans. on Information Theory, July 1971.

6. L. D. Davisson, "Data Compression - Theory and Application," Proc. University of Missouri Communications Conference, Oct. 1970.

7. T. J. Cruise, "Achievement of Rate-Distortion Bound Over Additive White Noise Channels Utilizing a Noiseless Feedback Channel," Proc. IEEE, vol. 55, 1967.

8. T. Kailath, "An Application of Shannon's Rate-Distortion Theory to Analog Communication over Feedback Channels," Proc. Princeton Conference on Systems, 1967.

9. J. P. M. Schalkwijk and L. J. Bluestein, "Transmission of Analog Waveforms Through Channels with Feedback," IEEE Trans. on Information Theory, July 1967.

10. L. D. Davisson, "Theory of Adaptive Data Compression," Advances in Communication, Academic Press, 1966.

11. T. J. Goblick, Jr., "Analog Source Digitization: A Comparison of Theory and Practice," IEEE Trans. on Information Theory, April 1967.

12. L. D. Davisson, "Comments on Sequence Time Coding for Data Compression," Proc. IEEE, Dec. 1966.

13. T. J. Lynch, "Sequence Time Coding for Data Compression," Proc. IEEE, Oct. 1966.

14. B. M. Fitingoff, "Optimal Coding for the Case of Unknown and Varying Signal Statistics," Problemy Peredachi Informatsi, vol. II, no. 2, 1966.

15. Y. M. Shtarkhov and B. F. Babkin, "Combinational Coding for Discrete Stationary Sources," Proc. of the International Symposium on Information Theory, Tsahkadsor, USSR, Sept. 1971.

ACKNOWLEDGEMENT

This research is supported by the National Science Foundation under Grant GK-141-90.

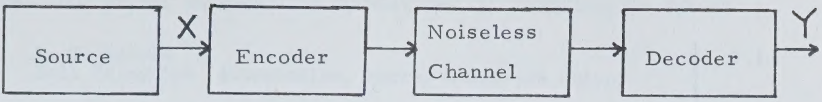

Source
X
Encoder
Noiseless
Channel
Decoder
Y

Figure 1. Data Compression System

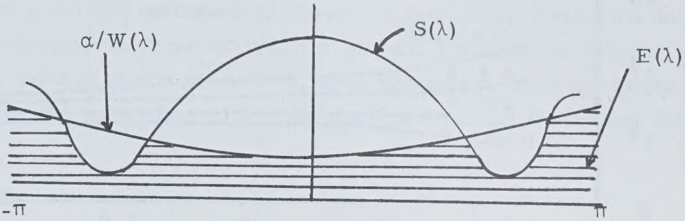


Figure 2. Illustration for Rate Distortion Function

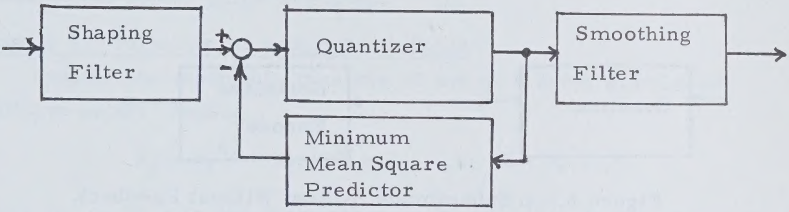

Shaping
Filter
+
-
Quantizer
Smoothing
Filter
Minimum
Mean Square
Predictor

Figure 3. A Data Compression System

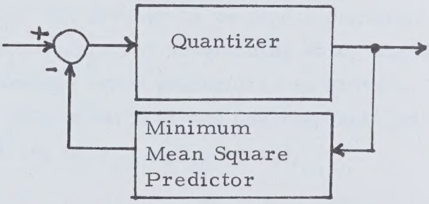

+
-
Quantizer
Minimum
Mean Square
Predictor

Figure 4. A Suboptimum System Using Feedback

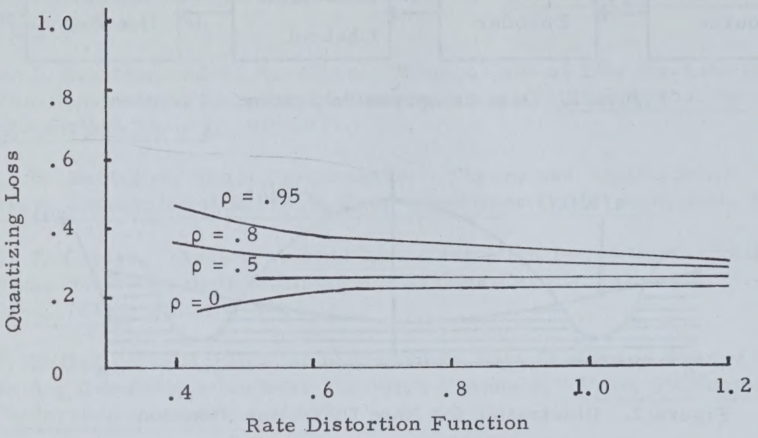


Figure 5. Bound on Quantizing Loss

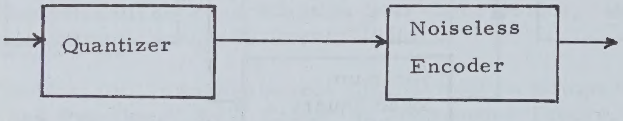

Quantizer
Noiseless
Encoder

Figure 6. A Suboptimum System Without Feedback

A NEW LOOK AT THE MATHEMATICAL FOUNDATIONS OF COMMUNICATION THEORY

T. T. KADOTA
Bell Telephone Laboratories, Murray Hill, New Jersey

ABSTRACT

We give a unified mathematical theory of detection, estimation and information theory, in terms of transfer probability measures and their derivatives defined on the product of input and output function spaces. We illustrate the main ideas with elementary probability theory and simple examples.

Introduction

Detection, estimation and information theories, three of the major branches of communication theory, have been developed more or less independently from each other. Consequently, their traditional mathematical formulations appear quite unrelated.

Example 1 Detection of m Signals in Noise

Suppose the observable consists of one of m known signals and additive noise. Namely,

$$y_i = s_1^{(k)} + z_i\ , \quad i = 1,\dots,n;\ \ k = 1,\dots,m, \tag{1}$$

where $(y_1,\dots,y_n)$ is the observable, $(s_1^{(k)},\dots,s_n^{(k)})$ is the signal, a sequence of real constants, with a priori probability $1/m$, and $(z_1,\dots,z_n)$ is the noise, a sequence of independent real Gaussian variables with mean 0 and variance 1. The problem is to find a decision rule, based on the observation of $y_1,\dots,y_n$, for determining which signal is present in such a way that the average error probability is minimum. Solution proceeds as follows: It is well known that the desired decision rule is to compare the m possible densities of $y_1,\dots,y_n$, namely,

$$p_n^{(k)}(y_1,\dots,y_n) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_1^n (y_i - s_i^{(k)})^2\right],\ k = 1,\dots,m, \tag{2}$$

and to choose the signal $s_1^{(k)},\dots,s_n^{(k)}$ for which the above density is the largest. When the bracket in the exponential is expanded, we note the first term $\sum_i y_i^2$ is independent of k. Suppose we impose the equi-power constraint on the signals, namely, $\frac{1}{n}\sum_i (s_i^{(k)})^2 = \rho$, $\rho > 0$, for all k. Then comparison of the m densities becomes equivalent to comparison of the middle term $\sum_i s_i^{(k)} y_i$. Thus, the desired rule is to choose the $s_1^{(k)},\dots,s_n^{(k)}$ for which $\sum_i s_i^{(k)} y_i$ is the largest.

Example 2 Estimation of Random Signal in Noise

Suppose the observable consists of a random signal and the same additive noise. Namely,

$$y_i = u_i + z_i , \quad i = 1,\ldots,n , \tag{3}$$

where the random signal $(u_1,\ldots,u_n)$ is a sequence of real Gaussian variables with mean 0 and covariance matrix (r_{ij}), and is independent of $(z_1,\ldots,z_n)$. The problem is to find an estimate $\hat{u}_n$ of u_n, based on the observation of $y_1,\ldots,y_n$, such that its mean-square error $E(u_n-\hat{u}_n)^2$ is minimum among all possible estimates. It is well known that the desired estimate in general is $E\{u_n|y_1,\ldots,y_n\}$, the conditional expectation of u_n given $y_1,\ldots,y_n$. Since $u_1,\ldots,u_n$ are Gaussian, hence $y_1,\ldots,y_n$ are Gaussian also, we have

$$E\{u_n|y_1,\ldots,y_n\} = \sum_{i=1}^{n} h_{ni} y_i \tag{3a}$$

where $h_{n1},\ldots,h_{nn}$ are constants and it is easy to show that they are the solution of the following system of equations:

$$h_{ni} + \sum_{j=1}^{n} h_{nj}\, r_{ji} = r_{ni}, \quad i = 1,\ldots,n . \tag{4}$$

Example 3 Coding Theorem of Information Theory

Consider the additive noise channel described by

$$y_i = x_i + z_i , \quad i = 1,2,\ldots, \tag{5}$$

where x_1, $x_2,\ldots$ and y_1, $y_2,\ldots$ are the input and the output of the channel respectively, and z_1, $z_2,\ldots$ are the same additive noise. Note $x_1,x_2,\ldots$ are real numbers, not random variables. A fundamental question of information theory is, under the average input power constraint, whether or not it is possible to transmit messages through this channel in such a way that error probability can be made arbitrarily small. The answer is in the affirmative if the rate of information transmission is below the capacity C of the channel. More explicitly, the coding theorem [1] asserts that given any $\varepsilon > 0$ and $R < C$, where $C = \frac{1}{2}\log(1+\rho)$ for this channel, there always exist some m input sequences (signals) of length N, $(x_1^{(k)},\ldots,x_N^{(k)})$, $k = 1,\ldots,m$, satisfying the average power constraint $\frac{1}{n}\sum_{i=1}^{n}[x_i^{(k)}]^2 \le \rho$ and a decision rule based on the observation of $y_1,\ldots,y_N$

(the observable), such that the error probability for each input is less than ε while the requirement on information rate $(\log m)/N \geq R$ is satisfied. The crucial condition is that the required information rate is less than the channel capacity, namely, $R < C$. Besides the channel of Example 3, the coding theorem holds for a general class of stationary memoryless channels and its generalization. [2]

Let us investigate in more detail the underlying general mathematical structure. The basic framework is the product space $X \times Y$ of sequences of real numbers $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots)$, labeled as the input and the output respectively of a channel. Note $y_1, y_2, \ldots$ are no longer random variables. The channel is specified by what we call an n-dimensional transfer probability density $q_n(y_1, \ldots, y_n; x_1, \ldots, x_n)$ defined on $X \times Y$ for every $n = 1, 2, \ldots$. q_n is the n-dimensional output probability density for each fixed input (the first n components), and for the channel of Example 3 it is

$$q_n(y_1, \ldots, y_n; x_1, \ldots, x_n) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \sum_{1}^{n} (y_i - x_i)^2\right] . \tag{6}$$

For any set B of sequences $(y_1, \ldots, y_n)$, define $Q_n(B, x)$ by

$$Q_n(B, x) = \int \cdots \int_B q_n(y_1, \ldots, y_n; x_1, \ldots, x_n) dy_1 \ldots dy_n \; . \tag{7}$$

For any m inputs $(x_1^{(k)}, \ldots, x_n^{(k)})$, $k = 1, \ldots, m$, an m-partition $(B_1, \ldots, B_m)$ of Y where B's are sets of $(y_1, \ldots, y_n)$ is a decision rule, with the interpretation that if the "observable" y belongs to B_k then we decide on the k-th "signal" $x^{(k)}$. The error probabilities associated with $x^{(k)}$, $k = 1, \ldots, m$, and $(B_1, \ldots, B_m)$ are given by $1 - Q_n(B_k, x^{(k)})$, $k = 1, \ldots, m$. By assigning an n-dimensional input probability density $p_{X,n}$ on X, we define $p_{X,n}$ on X, we define the joint density $p_{XY,n}$ of the input and the output by

$$p_{XY,n}(x_1, \ldots, x_n, y_1, \ldots, y_n) = p_{X,n}(x_1, \ldots, x_n) q_n(y_1, \ldots, y_n; x_1, \ldots, x_n) \; , \tag{8}$$

and the output density $p_{Y,n}$ as a marginal of $p_{XY,n}$. Denote by f_n the ratio of the joint density $p_{XY,n}$ to the product $p_{X,n} p_{Y,n}$ of the input and the output densities, namely,

$$f_n = p_{XY,n}/p_{X,n} p_{Y,n} = q_n/p_{Y,n} \; . \tag{9}$$

Then the mutual information between the input and the output, denoted by

$I_n(X,Y)$, is given by

$$I_n(X,Y) = \int_{-\infty}^{\infty} \cdots \int [\log f_n(x_1,\ldots,x_n,y_1,\ldots,y_n)]$$

$$\cdot p_{XY,n}(x_1,\ldots,x_n,y_1,\ldots,y_n)dx_1\ldots dx_n dy_1\ldots dy_n, \tag{10}$$

and the capacity C is given by

$$C = \lim_{n\to\infty} \sup_{p_{X,n}\in\rho_n} \frac{1}{n} I_n(X,Y) = \sup_{p_{X,1}\in\rho_1} I_1(X,Y) \tag{11}$$

for the stationary and memoryless channel. The class of ρ_n of $p_{X,n}$ over which the sup is taken, is determined by the expectation version of the constraint put on $(x_1^{(k)},\ldots,x_n^{(k)})$. For Example 3 where the average power constraint is used, ρ_n is the class of $p_{X,n}$ for which

$$\int_{-\infty}^{\infty} \cdots \int \frac{1}{n} \sum_{1}^{n} x_i^2 \, p_{X,n}(x_1,\ldots,x_n)dx_1\ldots dx_n \le \rho \tag{12}$$

Then the coding theorem asserts the existence of some m and N satisfying $(\log m)/N \ge R$, some $(x_1^{(k)},\ldots,x_N^{(k)})$, $k = 1,\ldots,m$, satisfying the given constraint and some m-partition $(B_1,\ldots,B_m)$ of Y, specifiable in terms of $y_1,\ldots,y_N$, such that $Q_N(B_k,x^{(k)}) \ge 1 - \varepsilon$, where ε and $R < C$ are given.

The detection problem illustrated by Example 1 is ambiguously formulated though such a formulation is customary. Mathematically, it makes no sense to say that a random variable y_i can be equal to any one of m random variables $s_i^{(k)} + z_i$, $k = 1,\ldots,m$. As the solution indicates, however, the decision is not on the signal but on the probability density of the observable, and (1) is merely used to specify the m possible densities of the random variables $y_1,\ldots,y_n$. Thus, restating the problem in the representation space R^n of $y_1,\ldots,y_n$, what is given are the m densities $p_n^{(k)}$, $k = 1,\ldots,m$, of (2) defined on R^n, and what is required is partitioning of R^n into m parts.

As a better alternative, let us formulate it in the framework of the product space of the input and the output of a channel, which we used for the information-theoretic problem. Interpret the known signals and the observable as the input and the output of a channel, which is specified by a transfer probability density q_n. The q_n in Example 1 is the same as the one in Example 3, namely, (6). Unlike the information-theoretic problem, the input probability is given as the a priori probability, and it is specified in terms of a discrete distribution, instead of a density, namely,

$P_{X,n}(\{x^{(k)}\}) = \alpha_k$ where $\sum_1^m \alpha_k = 1$ and $\{x^{(k)}\}$ denotes the set which contains a single element $x^{(k)}$. For any set B specifiable in terms of $y_1, \ldots, y_n$, where y_i is the i-th component of an element y of Y and is no longer a random variable, we define the joint probability $P_{XY,n}$ of the input and the output by

$$P_{XY,n}(\{x^{(k)}\} \times B) = P_{X,n}(\{x^{(k)}\}) Q_n(B, x^{(k)}) \tag{13}$$

where Q_n is defined by (7), and the output density $p_{Y,n}$ by

$$P_e = 1 - \sum_{k=1}^{m} P_{XY,n}(\{x^{(k)}\} \times B_k) = \sum_{k=1}^{m} P_{X,n}(\{x^{(k)}\})[1 - Q_n(B_k, x^{(k)})] \, . \tag{14}$$

The meaning of the word "average" becomes obvious when we recognize $1 - Q_n(B_k, x^{(k)})$ as the error probability for the k-th input. According to a fundamental theorem of detection theory, the partition which minimizes P_e is given by

$$B_k^* = \{y: \alpha_k f_n(x^{(k)}, y) > \alpha_j f_n(x^{(j)}, y) \text{ for every } j \neq k\} \, , \qquad k = 1, \ldots, m, \tag{15}$$

where f_n is given by the second expression of (9). If more than one f_n coincide on an element y, we assign that element to the B_k with the lowest corresponding subscript. The proof of this theorem will be given in the next section in a more general form.

In the estimation problem illustrated by Example 2, two sequences of random variables $(u_1, \ldots, u_n)$ and $(y_1, \ldots, y_n)$, related to each other by (3), are given and another random variable $\hat{u}_n$ is to be found. Clearly, the basic space is the so-called underlying abstract probability space $(\Omega, \mathfrak{F}, P)$ and u's and y's are functions of $\omega \in \Omega$. Although u's and y's can still be labeled as the signal and the observable, the physical background of measuring a random signal in a noisy environment is considerably obscured. In addition, the abstract ω-space is inconvenient to work with, and hence we need some representation space . As a special kind of representation space, let us use the input-output product space to reformulate the estimation problem. As in the case of the detection problem, we interpret the random signal and the observable as the input and the output of a channel, which is specified by q_n. Again the q_n for Example 2 is the same, namely (6). The input density $p_{X,n}$ is that of the random variables $u_1, \ldots, u_n$. Any estimate of the n-th component x_n of the input x, based on the observation of the first n components $y_1, \ldots, y_n$ of the output y, is a function g_n of $y_1, \ldots, y_n$. Its mean-square error $\overline{e^2}$ is given by

$$\overline{e^2} = \int_{-\infty}^{\infty} \cdots \int [x_n - g_n)y_1, \ldots, y_n)]^2$$

$$\cdot p_{XY,n}(x_1, \ldots, x_n, y_1, \ldots, y_n)dx_1 \ldots dx_n dy_1 \ldots dy_n \,, \tag{16}$$

where $p_{XY,n}$ is defined by (8). Then a fundamental theorem of estimation theory states that the estimate which minimizes $\overline{e^2}$ is given by

$$g_n^*(y_1, \ldots, y_n) = \int_{-\infty}^{\infty} \cdots \int x_n f_n(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

$$\cdot p_{X,n}(x_1, \ldots, x_n)dx_1 \ldots dx_n \,, \tag{17}$$

where f_n is defined by (9). That g_n^* corresponds to the conditional expectation $E\{u_n | y_1, \ldots, y_n\}$ of Example 2, may be easily seen with the aid of the following heuristic:

$$E\{x|y\} = \int x \; p(x|y)dx = \int x \, [p(x,y)/p(y)]dx = \int x \, [p(x,y)/p(x)p(y)]p(x)dx \,.$$

The precise statement of this fact as well as the proof of the theorem will be given in the next section in a more general form.

In summary, we have illustrated that the mathematical frameworks for the traditional formulation of the three problems are very different from each other. For the detection problem, it is the space of observables R^n and m probability densities defined on it. For the estimation problem, it is the abstract probability space $(\Omega, \mathfrak{F}, P)$ and two groups of functions defined on it. For the information-theoretic problem, it is the input-output product space and the transfer probability density defined on it. It is our opinion that the framework for the information-theoretic problem is most desirable since it reflects the physical background without compromising the mathematical rigor. Thus, we have reformulated the detection and the estimation problems in the framework used for the information-theoretic problem.

Finally, let us make a few general remarks. Whether it is a detection, estimation of information-theoretic problem, the channel is always given, independent of the nature of the problem on hand. Physically, the channel is a given medium to be used for communication or measurement. Mathematically, it is specified by a transfer probability defined on the product space of the input and the output of the channel. It describes probabilistically the output behavior of the channel in response to each input, just the way the transfer function of a linear system does deterministically. The usage of the channel is primarily specified by the input probability which we assign. This assignment varies a great deal from one problem to another. With the transfer probability given and the input probability assigned, the joint probability of the input and the output is immediately defined, and the output probability is given as one of its marginals. Thus we have the ratio of the joint to the product probabilities. Then, loosely speaking, all three problems can be stated in terms of the joint probability and their solutions are given in terms of this ratio.

In general, probabilities (transfer, input, etc.) are given as probability measures defined on σ-fields of sets, not through densities as they have been so far. This is specially true in the case where the input and the output spaces are function spaces instead of sequence spaces. In the next section we formulate the three problems in the product space of the input and the output functions, and provide three illustrative examples, which are the function-space analogues of Examples 1-3.

II. Unified Formulation in Function Space

Let X and Y be spaces of all real functions x and y on $[0,\infty)$. Denote by $\mathcal{A}_t$ the σ-field generated by the class of cylinder sets $\{x: a_i \leq x(t_i) < b_i, i = 1,\ldots,n\}$, $0 \leq t_i < t$, where a_i, b_i, t_i and n are arbitrary. Denote by $\mathcal{B}_t$ the σ-field of y sets similarly generated. Let $Q(\cdot,x)$ be a transitional probabilility measure defined on $\mathcal{B}_\infty$ for each $x \in X$. (cf. (7)). By definition $Q(B,\cdot)$ is $\mathcal{A}_\infty$ -measurable for every $B \in \mathcal{B}_\infty$[3]. (We need this measurability for defining the joint measure through integration by (16)). We refer Q as a transfer probability measure of a channel, and X and Y as its input and output spaces. We say Q is causal if $B \in \mathcal{B}_t$ implies $\mathcal{A}_t$-measurability of $Q(B,\cdot)$. Loosely speaking, this means that probabilities of output events do not depend on the future inputs.

Let $P_{X,t}$ be an input probability measure defined on $\mathcal{A}_t$. Denote by Q_t the restriction of Q to $\mathcal{B}_t$. Since Q is causal, Q_t is necessarily $\mathcal{A}_t$-measurable, and hence it is independent of $\{x(s), t \leq s < \infty\}$. Define the joint measure $P_{XY,t}$ of the input and the output on $\mathcal{A}_t \times \mathcal{B}_t$ by

$$P_{XY,t}(A \times B) = \int_A Q_t(B,x) P_{X,t}(dx) \tag{16}$$

for any $A \in \mathcal{A}_t$ and $B \in \mathcal{B}_t$. (cf. (8))[3]. Then the output measure $P_{Y,t}$ defined on $\mathcal{B}_t$ by $P_{Y,t}(B) = P_{XY,t}(X \times B)$, and the product measure $P_{X \times Y,t}$ is defined on $\mathcal{A}_t \times \mathcal{B}_t$ by $P_{X \times Y,t}(A \times B) = P_{X,t}(A) P_{Y,t}(B)$. [3] If $P_{XY,t}$ is absolutely continuous with respect to $P_{X \times Y,t}$, which we denote by $P_{XY,t} < P_{X \times Y,t}$, there is a Radon-Nikodym derivative $dP_{XY,t}/dP_{XY,t}$, which we denote by f_t . $P_{XY,t} < P_{X \times Y,t}$ also implies $Q_t < P_{Y,t}$ a.e. $[P_{X,t}]$ and

$$f_t = dP_{XY,t}/dP_{X \times Y,t} = dQ_t/dP_{Y,t} \quad \text{a.e. } [P_{X \times Y,t}] \tag{17}$$

where the second equality means that for almost every $x[P_{X,t}]$ the Radon-Nikodym derivative of $Q_t(\cdot, z)$ with respect to $P_{Y,t}$ exists and it is equal to f_t a.e. $[P_Y]$. (cf. (9)).

Detection Problem

Given a causal transfer measure Q of a channel and an input measure $P_{X,t}$ specified by $P_{X,t}(\{x^{(k)}\}) = \alpha_k$, $\sum_1^m \alpha_k = 1$, find a $\mathcal{B}_t$-measurable m-partition $(B_1^*, \ldots, B_m^*)$ of Y such that $\sum_1^m P_{XY}(\{x^{(k)}\} \times B_k^*) \geq \sum_1^m P_{XY}(\{x^{(k)}\} \times B_k)$ for any $\mathcal{B}_t$-measurable m-partition $(B_1, \ldots, B_m)$ of Y.

Note $\mathcal{B}_t$-measurability of $(B_1, \ldots, B_m)$ is the mathematical meaning of a decision rule being based on the observation of $\{y(s), 0 \leq s < t\}$. Also note $Q_t(\cdot, x^{(k)}) < P_{Y,t}$, $k = 1, \ldots, m$, and $P_{XY,t} < P_{X \times Y,t}$. Hence f_t exists.

Theorem 1

$$B_k^* = \{y: \alpha_k f_t(x^{(k)}, y) > \alpha_j f_j(x^{(j)}, y),\ j<k;\ \alpha_k f_t(x^{(k)}, y) \geq \alpha_j f_t(x^{(j)}, y),\ j>k\}$$

(cf. (15))

Proof

$$\sum_{k=1}^m P_{XY,t}(\{x^{(k)}\} \times B_k^*) - \sum_{k=1}^m P_{XY,t}(\{x^{(k)}\} \times B)$$

$$= \sum_{j,k=1}^m [\alpha_k\, Q_t(B_j B_k^*, x^{(k)}) - \alpha_j\, Q_t(B_j B_k^*, x^{(j)})]$$

$$= \sum_{j,k=1}^m \int_{B_j B_k^*} [\alpha_k f_t(x^{(k)}, y) - \alpha_j f_t(x^{(j)}, y)] P_{Y,t}(dy)$$

$$\geq 0 .$$

Estimation Problem

Given a causal transfer measure Q of a channel and a general input measure $P_{X,t}$ find a $\mathcal{B}_t$-measurable function g_t^* on Y such that

$$\int_{X \times Y} [x(t) - g_t^*(y)]^2 P_{XY,t}(d(x,y)) \leq \int_{X \times Y} [x(t) - g_t(y)]^2 P_{XY,t}(d(x,y))$$

for any $\mathcal{B}_t$-measurable function g_t on Y.

Theorem 2

If $P_{XY,t} < P_{X \times Y,t}$, then

$$g_t^*(y) = \int_X x(t)f_t(x,y)P_{X,t}(dx) \ . \qquad \text{(cf. (17))}$$

Proof

$$\int_{X\times Y} [x(t)-g_t(y)]^2 P_{XY,t}(d(x,y))$$

$$= \int_{X\times Y} ([x(t)-g_t^*(y)]+[g_t^*(y)-g_t(y)])^2 P_{XY,t}(d(x,y))$$

$$= \int_{X\times Y} [x(t)-g_t^*(y)]^2 P_{XY,t}(d(x,y)) + \int_{X\times Y} [g_t^*(t)-g_t(y)]^2 P_{XY,t}(d(x,y))$$

where we have used the following orthogonality property for the second equality: for any $\mathcal{B}_t$-measurable v_t

$$\int_{X\times Y} [x(t)-g_t^*(y)]v_t(y)P_{XY,t}(d(x,y)) = 0$$

since from the definitions of g_t^* and f_t

$$\int_{X\times Y} v_t(y)g_t^*(y)P_{XY,t}(d(x,y)) = \int_Y v_t(y)\int_X x(t)f_t(x,y)P_{X,t}(dx)P_{Y,t}(dy)$$

$$= \int_{X\times Y} x(t)v_t(y)P_{XY,t}(d(x,y)) \ .$$

Remark

g_t^* may be referred to as "the conditional expectation of $x(t)$ given $\{y(s),\ 0 \leq s < t\}$" in the sense that g_t^* is $\mathcal{B}_t$-measurable and for any $B \in \mathcal{B}_t$

$$\int_B g_t^*(y)P_{Y,t}(dy) = \int_B\int_X x(t)f_t(x,y)P_{X,t}(dx)P_{Y,t}(dy) = \int_{X\times B} X(t)P_{XY,t}(d(x,y)).$$

Information-theoretic Problem

Let Q be a causal, incrementally stationary and memoryless [4], transfer measure of a channel, and $\mathcal{P}_t$ the class of input measure $P_{X,t}$ on $\mathcal{Q}_t$ satisfying the input power constraint

$$\int_X \frac{1}{t}\int_0^t x^2(s)ds\ P_{X,t}(dx) \leq \rho \ . \qquad \text{(cf. (12))}$$

The mutual information $I_t(X,Y)$ between the input and the output is given by

$$I_t(X,Y) = \begin{cases} \int_{X\times Y} \log f_t(x,y)P_{XY,t}(d(x,y)) & \text{if } P_{XY,t} < P_{X\times Y,t} \\ \infty & \text{otherwise} \end{cases}$$

(cf. (10)), and the capacity of the channel with the average power constraint is given by

$$C = \overline{\lim_{t\to\infty}} \sup_{P_{X,t}\in\mathcal{P}_t(\rho)} \frac{1}{t} I_t(X,Y) \text{ . (cf. (11))}$$

Coding Theorem

Given $\varepsilon > 0$ and $R < C$, there exists some m and T satisfying $(\log m)/T \geq R$ and some $\{x^{(k)}(t),\ 0 \leq t < T\}$, $k = 1,\ldots,m$, satisfying $\frac{1}{T}\int_0^T [x^{(k)}(t)]^2 dt \leq \rho$ and some $\mathcal{B}_T$-measurable m-partition $(B_1,\ldots,B_m$ of Y such that $Q(B_k,x^{(k)}) \geq 1-\varepsilon$, $k = 1,\ldots,m$.

The proof is given in [4] .

Finally, we give three examples which are the function-space analogues (or the continuous-time analogues) of Examples 1-3. The purpose of these examples is to exhibit the concrete final results. Hence we shall only outline the derivation and refer the detailed steps to appropriate literature. The channel to be used is common and is described by

$$y(t) = \int_0^t x(s)ds + W(t), \quad 0 \leq t < \infty$$

where $\{W(t),\ 0 \leq t < \infty\}$ is a standard Wiener process. (cf. (5)). Then the transfer probability measure of this channel is the Gaussian measure with mean $\int_0^t x(s)ds$ and covariance $\min(t,t')$, $0 \leq t,t' < \infty$. (cf. (6))

Example 1

The input measure $P_{X,t}$ is given by $P_{X,t}(\{x^{(k)}\}) = 1/m$, $k = 1,\ldots,m$, where $\frac{1}{t}\int_0^t [x^{(k)}(s)]^2 ds = \rho$. Let μ_W be a standard Wiener measure on $\mathcal{B}_t$. Then $Q_t(\cdot,x^{(k)}) \sim \mu_W$, $k = 1,\ldots,m$, and $P_{Y,t} \sim \mu_W$, where "$\sim$" denotes mutual absolute continuity. [5] Hence $f_t(x^{(k)},y) = (dQ_t/dP_{Y,t})(x^{(k)},y) =$ $[(dQ_t/d\mu_W)(x^{(k)},y)]/[dP_{Y,t}/d\mu_W)(y)]$, and it is shown in [5] that $(dQ_t/d\mu_W)(x^{(k)},y) = \exp\{\int_0^t x^{(k)}(s)dy(s) - \frac{1}{2}\int_0^t [x^{(k)}(s)]^2 ds\}$ a.e. $[\mu]$. Since $dP_{Y,t}/d\mu_W$ and $\int_0^t [x^{(k)}(s)]^2 ds = \rho t$ are independent of k, comparison of $f_t(x^{(k)},\cdot)$ and comparison of $\int_0^t x^{(k)}(s)dy(s)$, $k = 1,\ldots,m$, determine the same y set. Hence

$$B_y^* = \{y\colon\ \zeta_k(y) > \zeta_j(y),\ j < k;\ \zeta_k(y) \geq \zeta_j(y),\ j > k\}$$

where $\zeta_k(y) = \int_0^t x^{(k)}(s)dy(s)$.

Example 2

The input measure $P_{X,t}$ is induced by a zero-mean Gaussian process with a continuous covariance $r(s,\tau)$, $0 \leq s,\tau < t$. Then $Q_t(\cdot,x) \sim \mu_W$ a.e. $[P_{X,t}]$ and $P_{Y,t} \sim \mu_W$, [5] and $dQ_t/dP_{Y,t} = (dQ_t/d\mu_W)/(dP_{Y,t}/d\mu_W)$ a.e. $[P_{XY,t}]$. Let $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\varphi_1,\varphi_2,\ldots$ be the eigenvalues and orthonormal eigenfunctions of the integral operator with the kernel r. Define $\xi_i(x) = \int_0^t \varphi_i(s)x(s)ds$ a.e. $[P_{X,t}]$, $\eta_i(y) = \int_0^t \varphi_i(s)dy(s)$ a.e. $[P_{Y,t}]$, $i = 1, 2, \ldots$. Then $\{\xi_i\}$ and $\{\eta_i\}$ are independent zero-mean Gaussian variables with variances $\{\lambda_i\}$ and $\{1\}$ relative to $P_{X,t}$ and $P_{Y,t}$ respectively. Moreover $x(s) = \sum_1^\infty \varepsilon_i\varphi_i(s)$ and $y(s) = y(s) = \sum_1^\infty \eta_i \int_0^s \varphi_i(s')ds'$, $0 \leq s \leq t$, where the convergence is in the mean relative to $P_{X,t}$ and $P_{Y,t}$ respectively. [7,pp.373-5], [5] It can be inferred from the results in [5] that

$$dQ_t/d\mu_W = \exp\left(\sum_1^\infty \xi_i\eta_i - \frac{1}{2}\sum_1^\infty \varepsilon_i^2\right) \quad \text{a.e. } [P_{X\times Y,t}] ,$$

$$dP_{Y,t}/d\mu_W = \exp\left[\frac{1}{2}\sum_1^\infty \eta_i^2\,\lambda_i/(1+\lambda_i) - \frac{1}{2}\log(1+\lambda_i)\right]. \quad \text{a.e. } [P_{Y,t}] .$$

Therefore, it follows by substitution that

$$\int_X x(t)f_t(x,y)P_{X,t}(dx) = \sum_1^\infty (\lambda_i/1+\lambda_i)(\eta_i\varphi_i)(t) = \int_0^t h(t,s)dy(s) \quad \text{a.e. } [P_{Y,t}]. \tag{18}$$

where $h(t,s) = \Sigma(\lambda_i/1+\lambda_i)\varphi_i(t)\varphi_i(s)$. (cf. (3A)) With the use of Mercer's expansion $r(t,s) = \Sigma\lambda_i\varphi_i(t)\varphi_i(s)$, it is easy to show that h is the unique solution of the following integral equation (cf. (4))

$$h(t,s) + \int_0^t r(t,\tau)h(\tau,s)d\tau = r(t,s), \quad 0 \leq s \leq t .$$

Example 3

For any $P_{X,t} \in \mathcal{P}_t$, it follows from a formula in [6] that

$$2I_t(X,Y) = \int_X \int_0^t x^2(s)ds\ P_{X,t}(dx) - \int_Y \int_0^t [\int_X x(s)f_s(x,y)P_{X,s}(dx)]^2 ds P_{Y,t}(dy)$$

which is obviously bounded by ρt. Choose $P_{X,t}$ to be the measure induced by a zero-mean Gaussian process with covariance $\rho\exp(-\alpha|s-\tau|)$, $\alpha > 0$, $0 \leq s,\tau \leq t$. Then $\lambda_1 < 1/\alpha$ [7,pp.99-101], and by using (18) and Mercer's expansion, the second term is seen to be bound by $\lambda_1 t/2$. Hence, as $\alpha \to \infty$, the first term remains ρt while the second vanishes. Therefore, sup $I_t(X,Y)/t = \rho/2$ and $C = \rho/2$. The coding theorem holds with this capacity.

REFERENCES

1. R. Ash, Information Theory, Interscience Publishers, New York, 1965, p.66.
2. J. Wolfowitz, Coding Theorem of Information Theory, Springer-V and Prentice-Hall, Englewood Cliff, N.J. 1964.
3. J. Neveu, Mathematical Foundations of the Calculus of Probability, Holden-Day, San Francisco, 1965, pp. 73-77.

4. T. T. Kadota, "Coding theorem and its converse for continuous, incrementally stationary channels with finite incremental memory", submitted for publication in the IEEE Trans. on Information Theory.

5. L. A. Shepp, "Radon-Nikodym derivatives of Gaus sian measures", Ann. Math. Stat., 37, 2, April 1966, pp. 321-354.

6. T. T. Kadota, M. Zakai and J. Ziv. "Mutual information of the white Gaussian channel with and without feedback", IEEE Trans. on Information Theory, IT-17, 4, July 1971, pp. 368-371.

7. W. B. Davenport and W. L. Root, Radom Signals and Noise, McGraw-Hill, New York, 1958.

ACHIEVEMENTS IN IMAGE PROCESSING

OLEH JOHN TRETIAK

Massachusetts Institute of Technology, Cambridge, Massachusetts

I INTRODUCTION

Image processing arose through the application of Communications and Control Theory to the design of imaging systems. I believe that the term was coined by D.H. Kelly in a pioneering study of optical filtering [Kelly '61]. An early application of these ideas was made by Elias et al. in 1952, who indicated the relevance of the Fourier Transform to the study of optical systems. In the initial applications of image processing, techniques were borrowed directly from the theory of time systems, but in the past decade the field has matured, and many procedures specific to the imaging area have been developed. In this paper, I would like to review the high points of these discoveries. I will be forced to leave many important results unmentioned, but I hope to present the flavor of the area, and show how this field differs from the traditional theory that is its parent.

The marriage of system theory and imaging systems with the midwifery of the computer art, has invigorated traditional areas such as television, optics, and photography, and also spawned three amorphous offspring which, for the lack of a better name, are called image enhancement, image coding, and pattern recognition. In this review, I will summarize techniques that fall most closely into the first of these three areas. I feel that image coding, the art of digitizing pictures for efficient storage and transmission, is too specialized for this review, and pattern recognition is outside the scope of what one often considers image processing. This review is, therefore, restricted to systems that produce an image as its final output. I will choose topics that, in my opinion, add insight to the area. Most of the results I will mention are theoretical, but this theory is quite close to the realms of application.

II THE FOURIER TRANSFORM AND CONVOLUTIONS IN TWO DIMENSIONS

The Fourier transform is perhaps the most powerful tool in the study of continuous systems. In image processing, however, it also has both physical and practical significance. Physical, because there are optical systems that evaluate it, and practical, because many computer processing techniques would not be feasible without it.

The two-dimensional Fourier transform may be defined by the integral:

$$G(\vec{w}) = \int d\vec{x}\, g(\vec{x})\, e^{-2\pi i(\vec{x},\vec{w})} \qquad (1)$$

We are representing the coordinate pair (x,y) by $\vec{r}$, and the transform coordinate pair (u,v) by the vector $\vec{w}$. In two dimensions, the Fourier transform has many interesting and valuable properties that have no one-dimensional analogs. The double integral may be thought of as an iterated integral,

$$G(u,v) = (\int dx\ (\int dy\, g(x,y)\ e^{-2\pi iyv})\ e^{-2\pi ixu}) \qquad (2)$$

and this property is used in the numerical evaluation of the Fourier trans-

---

This work was supported by the National Institutes of Health Grant 3T01 GM 15006-0352, and Grant 5P01 GM 14940-05.

form with the Fast Fourier Transform algorithm [Stockham '66]. This property is present only in rectangular coordinates. The Fourier transform in polar coordinates is given by

$$G(w,\phi) = \int_0^\infty r\,dr \int_0^{2\pi} d\theta\; g(r,\theta)\; e^{-2\pi i r w \cos(\theta-\phi)} \tag{3}$$

$$\text{where } x = r\cos\theta \qquad u = w\cos\phi$$

$$y = r\sin\theta \qquad v = w\sin\phi$$

The above form is very unattractive, since the kernel of the transform does not separate. The kernel may be expanded in terms of Bessel functions [Papoulis ' 68,p163]

$$G(w,\phi) = \sum_{k=-\infty}^{+\infty} e^{-ik\phi}\,(-i)^k \int_0^\infty r\,dr\,J_k(2\pi r w) \int_0^{2\pi} d\theta\; e^{ik\theta}\, g(r,\theta) \tag{4}$$

and, in spite of appearances, this may be viewed as a simplification of sorts, since it is the cascade of three operations:

1) A Fourier series expansion in theta,

2) A sequence of Hankel transforms

$$S_k(w) = 2\pi\,(-i)^k \int_0^\infty r\,dr\,J_k(2\pi r w)\, f_k(r), \tag{5}$$

3) A Fourier series in phi.

The Fourier series may be evaluated with fast algorithms, but the Hankel transforms are much more difficult to compute. This form of the Fourier transform is useful in cases when the function to be transformed has some special properties in polar coordinates, such as radial symmetry.

The one-dimensional scaling law

$$\text{F.T.}(f(ax)) = \frac{1}{|a|}\,F\!\left(\frac{w}{a}\right) \tag{6}$$

has a non-trivial generalization to two dimensions, which is

$$\text{F.T.}(f(A\vec{r})) = \frac{1}{|\det A|}\,F(A_t^{-1}\vec{w}) \tag{7}$$

In this equation, A is a general non-singular matrix.

Partial fraction expansions are a very useful tool in system theory. Unfortunately, since polynomials of two variables do not have a property analogous to the fundamental theorem of algebra, there is no similar expansion for two-dimensional rational functions. As a rule, Fourier transforms of two-dimensional functions seem to be hard to come by. A set of analytic transforms may be generated from one-dimensional functions. Since Fourier transform in rectangular coordinates can be written as an iterated integral with two kernels, if a function is separable

$$f(x,y) = f_1(x)f_2(y) \tag{8}$$

its Fourier transform is also separable, and is the product of the one-dimensional transforms of the two functions;

$$F(u,v) = F_1(u)F_2(v) \tag{9}$$

There are procedures for expanding arbitrary functions in a sum of separable functions [Tretiak & Huang '70].

The two-dimensional impulse function may formally be defined as the product of two one-dimensional impulse functions

$$\delta(x,y) = \delta(x)\delta(y) \tag{10}$$

and its Fourier transform is equal to one. The distribution

$$f(x,y) = \delta(x) \tag{11}$$

is also useful, and is called an impulse sheet. Its Fourier transform is, surprisingly enough,

$$\text{F.T.}(\delta(x)) = \delta(v) \tag{12}$$

Two-dimensional convolution is defined by the equation

$$f_3(\vec{r}) = \int d\vec{x} f_1(\vec{x}) f_2(\vec{r}-\vec{x}) \tag{13}$$

and it is diagonalized by a Fourier transform

$$F_3(\vec{w}) = F_2(\vec{w}) F_1(\vec{w}) \tag{14}$$

Convolutions may be performed with optical systems [Goodman '68], and with digital computers [Huang '67, Stockham '66].

III LINEAR TRANSFORMATIONS IN IMAGE RESTORATION

One of the set problems in Image Processing is illustrated in Fig. 1. We are presented with an image, $f_o(x,y)$, which is related to an "ideal" image by some transformation $f_o = T(f_i)$. It is almost always assumed that the transformation is linear. It is not very safe to suppose that the transformation is shift invariant: while, to a fair approximation, time is a universal, spatial coordinates of images bear little resemblance to those of the object. We therefore represent the transformation by the equation

$$f_o(\vec{x}) = \int d\vec{w}\ f_s(\vec{w}) k(\vec{x},\vec{w}) \tag{15}$$

It is natural to analyze the above equation by examining its spectrum of eigenvalues. If the equation has eigenvalues equal to zero, it is not possible to invert it. If it contains a complete set of characteristic functions, one may use them to evaluate the inverse. The above approach is, however, not very productive, because there are no general analytical techniques and numerical evaluation with the current power of digital computers is usually unfeasible. One is restricted to special cases, and the most important of these is when the transformation is a convolution

$$k(\vec{x},\vec{w}) = h(\vec{x}-\vec{w}) \tag{16}$$

There are some other simple cases that are often overlooked, such as

$$f_o(\vec{x}) = g(\vec{x}) f_s(\vec{x}) \tag{17}$$

and

$$f_o(\vec{x}) = f_s(\vec{u}(\vec{x})) \tag{18}$$

A system function defined by a finite concatenation of these three transforms can be analyzed, and some solutions of this form have been obtained [Robbins '70].

It has been suggested that if the impulse response of a system "varies slowly with space", it can be approximated as being shift-invariant over so called "isoplanatic patches", though in spite of some attempts [Robbins '70] this problem has not yet been resolved.

We note in passing that the impulse response for Side-Looking Radar, perhaps the most successful application of image processing, is not shift-invariant.

If the set of functions presented at the input of Fig. 1 do not contain functions that lie in the kernel of the transformation, it is possible to recover the input from the output even when the transformation is singular. A straightforward example of this occurs when the transformation is a convolution, the transform of the impulse response is zero for some set of spatial frequencies, and the transform of the input is also zero for these frequencies. A non-trivial instance of this is when the system function is bandlimited but the input function is nonzero only over a bounded, specified region [Frieden '67].

In summary, while the issues associated with convolutions are well understood, the more general problem of shift-variant transformation is as yet in an unsatisfactory state.

IV IMAGING SYSTEMS OF THE SAMPLING TYPE

In many cases, the observed data is measured over a space of lower dimension that that on which the original function is defined. We may, for example, wish to infer the structure of a three-dimensional object from photographs. More conventional forms of this problem arise when we wish to represent an image with a television signal, or with numerical values in a computer.

The problem of representing an image over a regular grid was investigated by Peterson and Middleton '62, who found that the Whittaker sampling theorem also applies to functions of many dimensions. They also found that low-pass filtering before sampling is optimal in the mean. The relation between television signals and the Fourier Components of an image was found by Mertz and Gray in 1934, who, to my knowledge, were the first to apply Fourier techniques to image analysis.

The problem of determining the internal structure of an object has recently received much attention. It is formulated in terms of a model, which may be formulated in any number of dimensions. Suppose we have a function of two dimensions, and we obtain a measurement that is equal to a family of line integrals through the function

$$g(\ell,\theta) = \int dt\ f(\ell\cos\theta - t\sin\theta, \ell\sin\theta + t\cos\theta) \tag{19}$$

It is easy to show that the one-dimensional Fourier transform of this function is equal to the two-dimensional transform of the original function evaluated along a line in the transform domain normal to the directions of the line integrals. This is illustrated in Fig. 2. [Bracewell '56]. By collecting a set of such views at different angles we can estimate the Fourier transform of the original function, and consequently determine the function itself.

The above technique has been proposed for radio astronomy [Bracewell '56], electron microscopy [DeRosier and Klug '68], and X-ray imaging [Tretiak et al. '69].

V OPTICS AND SYSTEM THEORY

One of the most elegant achievements in this field is the development of a systematic procedure for analyzing optical systems [Vander Lugt '66]. The model is indicated in Fig. 3. The optical system consists of three types of components: transparencies, lenses, and regions consisting of free space. The system is analyzed in terms of a scalar function, $g_f(x,y)$, which is the time Fourier transform of one of the rectangular components of the electromagnetic field defined on a plane normal to the optic axis. The three blocks have the following system functions:

Transparency:

$$g_f(x,y) = t(x,y)s_f(x,y) \tag{20}$$

Lens:

$$a_f(x,y) = r_f(x,y)e^{+2\pi i\frac{f}{c}d_f(x,y)} \quad \text{inside lens} \tag{21}$$
$$= 0 \quad \text{outside lens.}$$

where $d_f(x,y)$ is the optical path length through the lens at (x,y) at frequency f.

Free space:

$$T_f(x,y) = h_{D,f}(x,y) \otimes g_f(x,y)$$

$$H_{D,f}(u,v) = \begin{cases} e^{2\pi iD\sqrt{(f/c)^2 - u^2 - v^2}} & u^2 + v^2 < f^2/c^2 \\ e^{-2\pi D\sqrt{u^2 + v^2 - (f/c)^2}} & u^2 + v^2 > f^2/c^2 \end{cases} \tag{22}$$

Systems are usually analysed in terms of the so-called Fresnel approximation, in which the lens characteristic for an axially symmetric lens of focal length F is given by

$$e^{-(\frac{\pi i f}{c}\frac{(x^2+y^2)}{F})} \tag{23}$$

and the spatial frequency response of free space is approximated by

$$e^{-\pi i(\frac{Dc}{f})(u^2 + v^2)} \tag{24}$$

With this approximation, it is easy to show that the system of Fig. 4 obtains the Fourier transform of the transparency. This is an exact analog of a chirp filter.

Though the propagation of the field variable is linear, only intensities, which are quadratic in the field variable, can be measured at light frequencies. It should be clear that the determination of the input to an optical system from intensity measurements alone is not only non-trivial, but almost impossible.

The above model may be used to represent coherent radiation propagating through an optical system. Incoherent light must be represented by the mutual coherence function, a correlation function of the field variables [Born & Wolf p491], and the propagation of the mutual spectral density function, the temporal Fourier transform of the mutual coherence function, is quite similar to the laws of propagation of the field quantity. The mutual coherence function cannot be found directly from the image, since only intensity measurements are possible, and since many different mutual coherence functions produce the same intensity distribution, this presents serious difficulties in system analysis. One can only hope to analyze an object on the basis of image measurements if the illumination is incoherent [Born & Wolf, p483].

VI NONLINEARITIES

Just as imaging systems are usually shift-variant, it is seldom that one encounters a linear image forming system. Some of the nonlinearities and indeterminacies were described in the previous section. Interactions between energy and matter, such as the ones that occur in an illuminated object, are typically nonlinear functions of the object parameters, and image forming systems are not linear either.

The nonlinearities of recording media are usually modeled by assuming that they are memoryless. For example, if the energy distribution falling on an imaging device is $f_1(x,y)$, it is assumed that the signal from it is

$$f_2(x,y) = g(f_1 \otimes h) \tag{25}$$

This sort of nonlinearity causes no difficulties in computer processing of images, since one can easily invert them [Oppenheim et al. '68].

VII NOISE IN IMAGE PROCESSING

From the important role that linear transformations play in image processing transformations, it is not surprising that disturbances and measurement errors are usually modeled as additive noise, and least-squares systems are investigated. The formulations are no different than those in one dimension, but the solutions are typically simpler since causality is not an issue. [Peterson & Middleton '62, Helstrom '67, Slepian '67].

When an image is recorded at low light levels, the noise that arises from the image forming energy, be it light, electrons, or X-rays, is usually governed by a Poisson process. Though this aspect of the noise has been considered in an ad-hoc way [Pizer & Vetter '70], analytic treatments of this process are lacking.

VIII CONCLUSIONS AND FORECASTS

By and large, the kind of image processing that we have discussed in this paper has not lived up to its promise a decade ago. In spite of a large amount of effort in image enhancement, best images are still to be obtained with good components and cannot be improved with image processing. The large amount of effort devoted to overcoming degradations due to atmospheric turbulence has shown meager results. The sort of processing that was carried on with pictures from space has yet to be brought down to earth.

The noteworthy results that have occurred seem to come not so much from improving on established ways of doing things, but rather with systems that use a novel method or procedure. This, for example, is the case with holography, Side-Looking Radar, and three-dimensional reconstruction procedures.

The state of the applications that are made at any time is limited by the general technological level. If computers come down in price, or if better optical processing methods are developed, it may be possible that the image processing procedures will gain wider use. We must also face the possibility that the techniques that have been applied to date are getting mined out. In spite of continuing gradual progress in the elaboration of the technology, the basic communications theory tools have been developed in the 1940-1950 decade, and the computer and the laser are novelties no longer. Perhaps the current founderings in pattern recognition and in computer science are laying the groundwork for another quantum increase in our understanding.

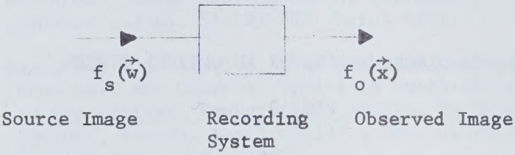
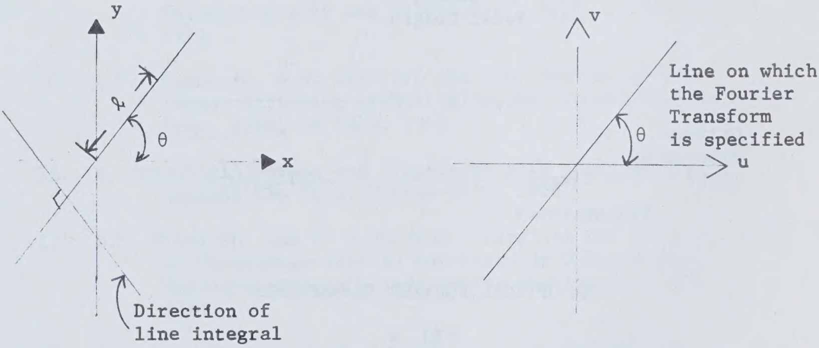
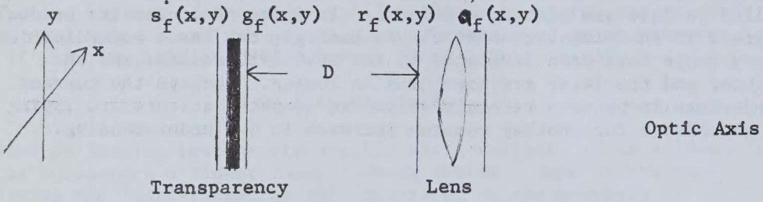

$f_s(\vec{w})$
$f_o(\vec{x})$
Source Image
Recording System
Observed Image

FIG. 1


y
x
θ
Direction of line integral
v
u
θ
Line on which the Fourier Transform is specified

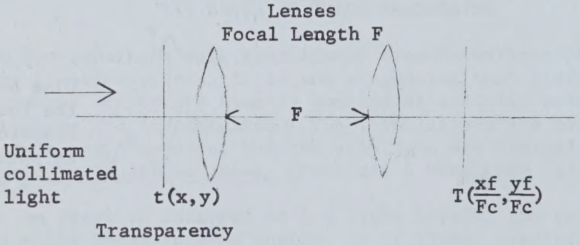
THE PROJECTION THEOREM

FIG. 2


y
x
$s_f(x,y)$
$g_f(x,y)$
D
$r_f(x,y)$
$q_f(x,y)$
Optic Axis
Transparency
Lens

BLOCK DIAGRAM MODEL OF AN OPTICAL SYSTEM

FIG. 3


Lenses
Focal Length F
F
Uniform
collimated
light
$t(x,y)$
$T(\frac{xf}{Fc},\frac{yf}{Fc})$
Transparency

AN OPTICAL FOURIER TRANSFORMER

FIG. 4

1. R.N. Bracewell, "Strip Integration in Radio Astronomy", Austr. J. of Phys., 9:198, 1956.

2. M. Born and E. Wolf, "Principles of Optics", London, Pergamon Press, 1959.

3. B.R. Frieden, "Band Unlimited Reconstruction of Optical Objects and Spectra", JOSA, 57:1013-1019, 1957.

4. P. Elias, D.S. Grey, and D.Z. Robinson, "Fourier Treatment of Optical Processes", JOSA, 42:127-134, 1952.

5. J.W. Goodman, Introduction to Fourier Optics, New York, McGraw-Hill, 1968.

6. C.W. Helstrom, "Image Restoration by the Method of Least Squares", JOSA, 57:297-303, March 1967.

7. T.S. Huang and H.L. Kasmitz, "The Combined Use of Digital Computers and Coherent Optics in Image Processing", in Proceedings, Computerized Imaging Techniques Seminar, Renondo Beach, Calif., Soc. of Photo-Optical Inst. Eng., 1967.

8. D.H. Kelly, "Image Processing Experiments", JOSA, 51:1095-1101, 1961.

9. P. Mertz and F. Gray, "A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television", BSTJ, 13:464-515, 1934.

10. A.V. Oppenheim, R.W. Schafer, and T.G. Stockham, "Nonlinear Filtering of Multiplied and Convolved Signals", Proc. IEEE, 56:1264, 1968.

11. A. Papoulis, Systems and Transforms with Applications in Optics, New York, McGraw-Hill, 1968.

12. D.P. Peterson, and D. Middleton, "Sampling and Reconstruction of Wavenumber-Limited Functions in N-Dimensional Euclidean Spaces", Information and Control, 5:279, 1962.

13. Pizer and Vetter, in B. Lipkin and A. Rosenfeld, eds., Picture Processing and Psychopictorics, New York, Academic Press, 1970.

14. D. Slepian, "Linear Least-Square Filtering of Distorted Images", JOSA, 57:918-922, July 1967.

15. T.G. Stockham, Jr., "High-Speed Convolution and Correlation", Spring, Joint Computer, Conference, AFIPS Proc., 28:229-233, 1966.

16. O.J. Tretiak, M. Eden, and W. Simon, "Internal Structure from X-ray Images", Proc. 8th Int. Conf. on Med. and Bio. Eng., Chicago, 1969.

17. O.J. Tretiak and T.S. Huang, "Separable-Sum Approximation of Two-Dimensional Digital Filters", Proceedings of the U.M.R. - Mervin J. Kelly Communications Conference, Rolla, Mo., University of Mo., Rolla, 1970.

18. A. Vander Lugt, "Operational Notation for the Analysis and Synthesis of Optical Data-Processing Systems", Proc. IEEE, 54:1055-1063, August 1966.

NOISELESS CODING FOR SIGNAL DESIGN

P. A. FRANASZEK
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

Abstract: A number of techniques are reviewed for the generation of codes used for matching system or transmission channel characteristics. Such constrained codes find application in a wide variety of digital transmission and recording systems for purposes including spectral shaping, reduction of intersymbol interference, and self-clocking.

I. Introduction

This paper discusses a number of techniques and problems associated with the coding of digital information to match transmission channel or system characteristics. Typical applications include spectral shaping, provision of a sufficiency of pulses or transitions between signal levels for purposes of clock recovery, and the suppression of intersymbol interference. Such criteria may often be best met by suitably encoding the digital information to be transmitted, rather than by the design of an appropriate modulation method or pulse shape.

Noiseless or constrained coding has been applied to a wide variety of systems utilizing bandwidth or input restricted channels. Typical examples are systems such as the AT&T T1 and T2, and modem transmission over telephone lines, where bandwidth restrictions limit the transmission rate. Examples of input restricted channels are encountered in digital magnetic recording on disks and tapes, and systems where certain distinct pulse sequences are reserved for such purposes as control and synchronization.

The codes under discussion may be considered as mappings from random or unformatted data to sequences which satisfy the desired constraints. Shannon [1] obtained an expression for the maximum information rate for an input restricted channel where the restrictions may be described in terms of a finite-state machine, a property which appears to be satisfied by constraints encountered in practice. Let S denote the set of channel states. To each $\sigma_i \in S$, there corresponds a set of allowable channel symbols. The transmission of a symbol takes the channel to a new state which is a function of the previous state and the transmitted symbol. A channel skeleton transmission matrix may be defined as follows: $D = [d_{ij}]$ where d_{ij} is the number of distinct channel symbols which take the channel from σ_i to σ_j.

The channel capacity C, defined as the number of bits per symbol that can be carried by the channel, may be expressed as the logarithm to the base two of the largest real root of

$$\det [d_{ij} Z^{-1} - \delta_{ij}] = 0$$

where δ_{ij} is the Kronecker delta.

Codes for the constrained channel may be linear or nonlinear. The latter generally employ table look up encoding. An exception is a class of codes where channel words are generated from blocks of binary digits via algebraic means. Linear coding is performed by passing the information sequence through a digital filter. Codes may be of fixed or of variable length, but appear to have the common property that the bit per symbol ratio is constant to avoid the need for buffers. In the following development, it will be seen that the notion of a sequence or channel state is a thread linking many of the techniques.

II. Nonlinear Codes

Consider the input restricted channel defined above. A code may be considered as a mapping $W_b^n \rightarrow \boxed{F(\sigma)} \rightarrow W_c^m$ where σ is the channel state, W_b^n is a binary n-tuple, and W_c^m is an m-tuple of channel symbols.

Freiman and Wyner[2] were among the earliest to consider systematic methods for constructing state oriented codes. They developed maximum block codes with the property that words may be freely concatenated without violating the channel constraints. That is, their codes are of the form

$W_b^n \rightarrow \boxed{F} \rightarrow W_c^m$ The method is particularly suited to finite memory channels, that is, channels in which knowledge of a fixed number of immediately previous inputs is always sufficient to uniquely specify the present channel state.

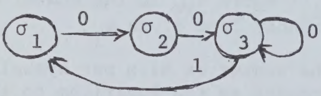
A set of words Q may be freely concatenated if each is allowable from all members of a set of states S' and takes the channel to a terminal state $\sigma(\omega) \in S'$. The method is based on finding terminal states associated with the maximal code word set.

Such state-independent codes have the advantage of lack of error propagation from word to word and freedom from having to track the channel state during the coding process.

Franaszek[3,4,5] studied techniques for finding the shortest codes with a given fixed bit per symbol ratio where there are no restrictions on state dependence. An additional introduced property was that the codes may be of variable length, although the bit per symbol ratio is constant.

The specification of the desired bit per symbol ratio permits a significant decrease in the computational complexity of the search for a code. The method is based on determining a set of states S' with the property that from each $\sigma \in$ S' there exist a sufficient number of distinct paths (terminating in other states belonging to S') to support the required bit per symbol ratio. The procedure for finding a set of such states is rather simple for fixed length codes, but requires the recursive application of a dynamic programing algorithm for the variable length case.

The following example[4] illustrates some of the advantages of variable length encoding. Consider the input restricted channel whose state transition diagram is illustrated below:



The restrictions require that each transmitted 1 be separated by at least two 0's, as might occur in a digital magnetic recorder where intersymbol interference is controlled by requiring that each transition between saturation levels be separated by at least two band intervals.

The capacity of the channel is approximately .55 bits/digit. The shortest fixed length code with a rate (bit per symbol ratio) of 1/2 is a state-independent code with a block length of 14. That is, the dictionary has a total of 128 words. The optimal variable length code is also state independent, with a dictionary of 3 words. The code is as follows:

Binary Block	Code Word
0	00
10	0100
11	1000

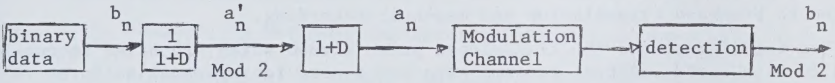
Although state dependent and variable length coding can result in significant savings in table look up requirements, the problem of error propagation is introduced. For the fixed length case, methods[3] have been developed for coding so that decoding is performed without knowledge of the channel state, although the coding process may be state dependent. This is essentially a nonlinear generalization of the precoding process in the partial response or correlative level coding discussed below. Error propagation is more complex for variable length codes, since an error in detection may affect the apparent word length at the decoder (misframing)[6]. For this case, techniques have been developed to generate optimal error propagation limiting codes for channels with finite memory[5].

A form of encoding which differs from the above methods was developed by Kautz[7]. The codes are binary, with the property that two consecutive ones are separated by no more than a specified number of zeros. Code word generation is performed by representing a binary information block in a form utilizing generalized Fibonacci weights. Tang and Bahl[8,9] extended the method to the care of binary sequences where two consecutive zeros are separated by at least d but no more than K zeros. Such codes have the drawback that words may not be freely concatenated without violating the imposed requirements. This may be countered by inserting a buffer sequence[9] between adjacent words. Since the length of the buffer sequences are only a function of the parameters d and K, the degradation in overall coding efficiency can be limited by increasing the code word length. This, however, frequently results in implementations which are more complex than those obtained from the state oriented methods discussed above.

III. Linear Encoding

In linear encoding, often termed partial response or correlative level coding, the information sequence is passed through a digital filter which provides a convenient method for spectral shaping at the expense of having to detect a greater number of levels.

A drawback of such systems is that errors in detection will tend to propagate. A method to avoid this problem was introduced by Lender[10] in connection with his Duobinary code. Let D represent $e^{-j\omega nT}$. A system block diagram is shown below



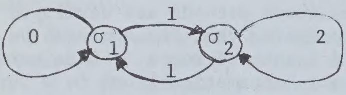
It can be noted that precoding does not affect the spectral properties of the system for statistically independent sequences $\{b_n\}$.

Numerous investigators including Van Gerwen[11], Kretzmer[12], and others [13, 14] have studied linear encoding methods of the above form. Such techniques are employed in a number of applications, including high performance data modems.

Some of the advantages of linear encoding are illustrated by the following example. Consider Lender's Duobinary. Croisier[15] notes that the spectral shaping introduced by the digital filter is such that the absolute value of the resulting pulse shape decays asymptotically as $^1/t^2$. Practically, this means that it is possible to transmit duobinary sequences in the Nyquist bandwidth. Uncoded binary pulses decay as 1/t. This implies that binary signaling may result in very large values between sampling points for unfavorable sequences. As a result, such signals require more bandwidth; in practice excess bandwidths of about 25 percent have been achieved, for example, with echo modulation techniques[16,17].

The above linear methods are in structure quite similar to convolutional encoding. Precoding to eliminate error propagation results in codes analogous to systematic convolutional codes.

The filter F(D) may also be regarded as a finite state machine producing outputs suited to a constrained channel of the type discussed above. For example, in the Duobinary technique, the output of the digital filter may be considered an input into a constrained channel with the state transition diagram shown below:



States σ_1 and σ_2 correspond to the previous transmission of a 1 and a 0 respectively. Precoding may then be considered a special case of coding so that one may decode independently of the channel state. That is, that a channel output of 1 uniquely determines the value of the transmitted information bit.

The state description of linear encoding permits one to combine some of the advantages of linear and nonlinear encoding methods. Desired spectral nulls may be introduced by choosing constraints corresponding to a linear filter, and additional restrictions imposed to satisfy such constraints as means for self-clocking. That is, one may consider the class of sequences that are possible outputs of such a linear filter, and choose a subset that satisfies the added requirements. Nonlinear codes of fixed or variable length may be constructed which produce sequences which are members of this subset[18].

The state description of the above encoding process is also useful in discussing means for maximum likelihood detection, which will be discussed briefly below.

IV. Some Applications

This section is a review of a few of the techniques which have been applied to baseband transmission and magnetic recording.

Two requirements on the transmitted pulse stream which are often encountered in high speed baseband transmission are a need for spectral nulls at d.c. and a high density of transitions. The first requirement is imposed by the necessity for a.c. coupling in most such systems. This constraint on the spectrum may be satisfied by linear codes, as well as such codes as bipolar[19], the latter being a code introduced by AT&T for the transmission of telephone signals by pulse code modulation on cables. These codes, however, are not suitable for systems with stringent timing requirements. Here the most common approach has been to use nonlinear table look up encoding. Although it is not easy to relate nonlinear sequence constraints to spectral properties, a convenient measure which appears to give a useful criterion for the amount of a.c. coupling degradation is the running digital sum variation (DSV)[3,15] defined as the difference between the upper and lower bounds of $\sigma_N = \sum_{n=M}^{N} a_n$ where a_n is the normalized d.c. contribution of the n^{th} transmitted digit. Constraints on the information rate imposed by a given bound on the DSV were studied by McCullough[19] and Chien[20].

The common approach to such coding is a compromise between code complexity, transition density, and DSV variation. References[21,22,15,23] give some of the better known codes. A particularly interesting one is the B6ZS of Johannes et al.[22] which has been adopted for use on the Bell System T2 6.3 megabit line[24]. Coding is done as in bipolar, with strings of six consecutive zeros replaced by special words. At first glance, it appears not to fit within the framework of Section II, but it is actually a state dependent variable length code.

Saturated digital magnetic recording is an example of an application where the channel itself imposes a set of constraints. The information is coded onto the location of transitions between saturation levels. Here, as in baseband transmission, timing requirements are stringent. Most codes adopted in practice fit within the class of binary codes with upper and lower limits on distances between transitions[25,26,27,5,7,8]. Codes are generally nonlinear, although there has been some study of the application of linear methods[28].

V. Concluding Remarks

Some of the approaches to coding for signal design were reviewed. It was seen that the notion of sequence or channel states is a thread connecting the area. This carries over to the problem of the optimal detection of a sequence that has been encoded for transmission. Here the work of Chang and Hanckok[29], Viterbi[30] and others including [31,32] have the common basis of the use of a state description of the interdependence of consecutively received symbols to formulate computationally feasible methods for maximum likelihood detection (MLD). Recently, Kobayashi and Tang[33] and Forney[34] applied these techniques to the problem of the MLD of correlatively encoded data. They showed that what was considered to be a penalty in signal to noise ratio for Duobinary versus binary transmission is really due to the nonoptimality of earlier detection schemes. This raises new questions concerning the advantages and costs of coding for signal shaping.

References

1. C. E. Shannon, "A Mathematical Theory of Communication," BSTJ, Vol. 27, pp. 379-423, 1948
2. C. V. Freiman and A. D. Wyner, "Optimum Block Codes for Noiseless Input Restricted Channels," Info. and Control, Vol. 7, pp. 398-415, 1964
3. P. A. Franaszek, "Sequence-state Coding for Digital Transmission," BSTJ, Vol. 47, pp. 143-157, 1968
4. P. A. Franaszek, "On Synchronous Variable Length Coding for Discrete Noiseless Channels," Info. and Control, Vol. 15, pp. 155-164, 1969
5. P. A. Franaszek, "Sequence-State Methods for Run-Length-Limited Coding," IBM J. of Res. & Development, Vol. 14, No. 4, pp. 376-387, July 1970
6. E. N. Gilbert and E. F. Moore, "Variable-Length Binary Encodings," BSTJ, Vol. 38, pp. 933-967, 1959
7. W. H. Kautz, "Fibonacci Codes for Synchronization Control," IEEE Trans. Information Theory, Vol. IT-11, pp. 284-292, April 1965

8. D. T. Tang, "Run-Length Limited Codes," 1969 IEEE International Symposium on Information Theory.
9. D. T. Tang and L. R. Bahl, "Block Codes for a Class of Constrained Noiseless Channels," Info. and Control, Vol. 17, No. 5, pp. 436-461, 1970
10. A. Lender, "The Duobinary Technique for High Speed Data Transmission," IEEE Trans. on Comm. & Electronics, Vol. 82, pp. 214-218
11. P. J. Van Gerwen, "On the Generation and Application of Pseudo-Binary Codes in Pulse Transmission," Phillips Research Reports, Vol. 20, 1965
12. E. R. Kretzmer, "Generalization of a Technique for Binary Data Communications," IEEE Trans. on Comm. Tech., Vol. COM-14, pp. 67-68, 1966
13. A. M. Gerrish, R. D. Howson, "Multilevel Partial Response Signaling," International Conference on Communications, Minneapolis, Minn., June 1967
14. F. K. Becker, E. R. Kretzmer, and J. R. Sheehan, "A New Signal Format for Efficient Data Transmission," BSTJ, Vol. 45, pp. 755-758, 1966
15. A. Croisier, "Introduction to Pseudoternary Transmission Codes," IBM J. of Res. & Dev., Vol. 14, pp. 354-367, 1970
16. M. Humbert, "4-Phase Digital Echo Modulation Technique," CCITT COM Sp. A Contribution Nr. 150-E. (Oct. 30, 1967)
17. A. Croisier and J. M. Pieret, "High Efficiency Data Transmission Through Digital Echo Modulation," International Communication Conference (1969) Boulder Col.
18. P. A. Franaszek, "Signaling Through a Partial Response Channel," Bell Laboratories Internal Memorandum, 1968
19. R. H. McCullough, "Ternary Codes for Regenerative Digital Transmission Systems," Ph.D. Dissertation, Polytechnic Institute of Brooklyn, June 1967
20. T. M. Chien, "Upper Bounds on the Efficiency of dc-Constrained Codes," BSTJ, Vol. 49, pp. 2267-2287, November 1970
21. J. M. Sipress, "A New Class of Selected Ternary Pulse Transmission Plans for Digital Transmission Lines," IEEE Trans. Comm. Tech., COM-13, No. 3, September 1965, pp. 366-372
22. V. I. Johannes, A. G. Kaim and T. Walzman, "Bipolar Pulse Transmission with Zero Extraction," IEEE Trans. Comm. Tech. COMj13, 303 (1969)
23. H. Kobayashi, "A Survey of Coding Schemes for Digital Data," to appear in IEEE Trans. Comm. Tech.
24. J. H. Davis, "T2: A 6.3 Mb/S Digital Repeatered Line," Int. Comm. Conf. (1969) Boulder, Colo.
25. C. M. Melas, "Quelques Proprietes des Codes Binaires avec Symbols Consecutifs," IEEE international Symposium on Information Theory, 1967
26. A. Gabor, "Adaptive Coding for Self-Clocking Recording," IEEE Trans. Elect. Computers, EC-16, 866 (1967)
27. D. T. Tang, "Practical Coding Schemes with Run-Length Constraints," IBM Research Report RC-2022, 1968
28. H. Kobayashi and D. T. Tang, "Application of Partial Response Channel Coding to Magnetic Recording Systems," IBM J. Res. & Dev., Vol. 14, No. 4, pp. 368-375, July 1970
29. R. W. Chang and J. C. Hanckok, "On Receiver Structures for Channels Having Memory," IEEE Trans. Inf. Theory, Vol. IT-12, pp. 463-468, October 1966
30. A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymtotically Optimum Decoding Algorithm," IEEE Trans. Information Theory, Vol. IT-13, pp. 260-269, April 1967
31. T. K. Onuera, "On the Viterbi Decoding Algorithm," IEEE Trans. on Information Theory, Vol. IT-15, pp. 177-179, January 1969

32. K. Abend et al., "On Optimum Receivers for Channels Having Memory," IEEE Trans. on Information Theory, Vol. IT-14, No. 6, Correspondence, pp. 819-820, November 1968
33. H. Kobayashi and D. T. Tang, "On Decoding and Error Control for Correlative Level Coding System," 1970 International Symposium on Information Theory, Abstract pp. 43-45
34. G. D. Forney, Jr., "Error Correction for Partial-Response Modems," 1970, International Symposium on Information Theory, June 1970

"COMPUTER REPRESENTATION AND MANIPULATION OF TRANSPORTATION NETWORKS"

VICTOR J. MARMA
Transportation Systems Center - Department of Transportation
55 Broadway, Cambridge, Massachusetts 02142

ABSTRACT

The concept of a network is basic to many forms of transportation systems. For advanced planning and research it is desirable to evaluate many alternative network configurations. The analysis is often done with the aid of computer programs for simulation and/or optimization; hence, it would be most advantageous to have a facility to interact with the user to easily effect changes in the network structures that serve as input to the simulation and/or optimization programs. This paper presents a methodology for the computer representation and manipulation of network structures.

Introduction

OPERATIONAL CONCEPTS

The object of central concern for our analysis is a computer capability to restructure networks. The box labelled "Network Manipulator" in Figure 1 represents such a capability.

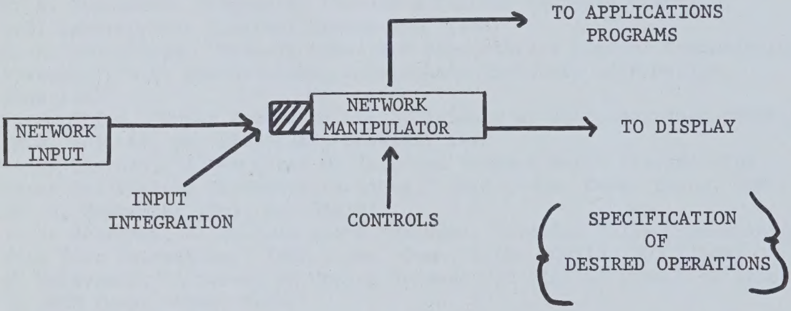


Figure 1

The main functional units of a system such as that pictured in Figure 1 are:

1) a set of simulation/optimization programs that have as input a network structure - the programs will concern such questions as network flow, traffic control and routing procedures,
2) the network manipulator - a software capability that can modify network structures and data,
3) set of controls - a specification of the possible operations for the network manipulator, and
4) a display system - a capability such that the user can visually see the network structures.

It is assumed that an average user will come to the facility with only a network map and some general questions. The specific capabilities of the library of simulation/optimization programs will be described to him as well the capabilities of the display - manipulator input system.

[The user may of course bring his own simulation/optimization programs.]

The facility will be used in an interactive iterative fashion. After the first run of a selected program and a given network configuration, results of the run together with the network will be displayed to the user. He then has the capability of changing the network structure and data before the next computer run.

This paper is primarily concerned with (2) and (3) above. The three basic parts of the analysis are i) the defining of a data structure for the representation of a network, ii) the developing of a symbolic language for operating on the data structure, and iii) the constructing of a user language i.e. a set of algorithms in terms of the symbolic language that are the input controls to the network manipulator.

Design Philosophy

The user will supply as input to the system a set of nodes and the connection links between them. For the initial analysis there will be only one type of connection possible between any pair of nodes, however the specification of different subsets of nodes will be allowed. It is considered that relatively minor additions will be used to encompass the possibility of different types of connections. The capabilities supplied to the user will be the abilities to alter the basic set of nodes and connections by:

1) adding nodes,
2) removing nodes,
3) adding connections,
4) removing connections.

In addition, the user will have the capability of constructing subnetwork structures from the total set of possibilities. In fact, the heart of the analysis is the development of the data structure, language and algorithms that can be used for the construction and manipulation of network structures.

In Figures 2 to 9 it is illustrated how a network substructure might be built. The detail step by step use of the USER language is not given here but will be presented in Section 2.5. In general, there will be a set of important nodes used as entry points to the display system (Figure 2). The user then selects the particular entry points of interest (Figure 3) and adds and/or subtracts nodes in the area of interest until a satisfactory pattern is obtained (Figures 4,5, & 6). The connections between the nodes are then shown (Figure 7). The next option is then that of modifying existing linkages until the desired network is obtained (Figures 8 & 9).

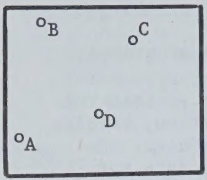


Figure 2

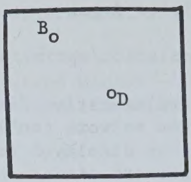


Figure 3

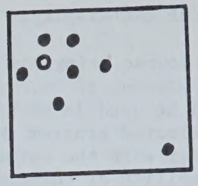


Figure 4

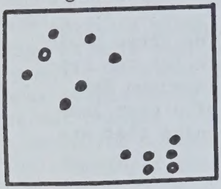


Figure 5

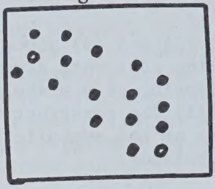


Figure 6

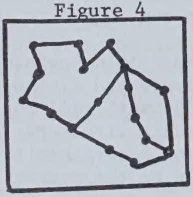


Figure 7

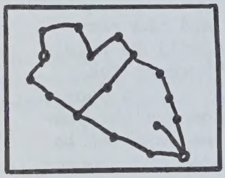


Figure 8

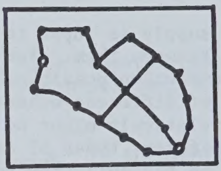


Figure 9

The interdependence of data structure, language, and algorithms must be recognized. It is only for purposes of presentation that we divide our analysis into sections but the actual work was not segmented as such.

All three aspects must be considered as a single system as they are interrelated. This is just a special case of the general idea that one's language or notation restricts the concepts or operations that are possible to consider. For example, it is almost impossible to translate certain concepts from one culture to another because of the different notation (set of categories) for viewing the world. For example, the Hopi language contains no words, grammatical forms, constructions or expressions that refer directly to what we call "time" or to past, present, or future or to enduring or lasting or to motion as kinematic rather than dynamic (i.e., dynamic refering to effort or expenditure of energy). Organizing experience in a fashion without our usual time concept is beyond our comprehension. We consider a particular network view and develop a procedure for operation within the confines of this view.

Organization of Paper

The main body of the paper is contained in SECTION 2. Analysis. The analysis section consists of five main subsections. The first defines the basic concept of representation. The second concerns the organization of the input data and the specification of the internal data structure. The third is concerned with the development of a notation or language to facilitate operations on the data structure. The fourth describes the operations performed on the data structures. Finally, the last subsection gives a listing of the commands as it would appear to the user.

ANALYSIS

Concept of Representation

It is well recognized that for the computer representation of graphs or networks that are subject to updating some kind of multi-list structure is most desirable. The basic approach to data has been to form a list structure representing a network. When modifications are desired in the network a change is made in the corresponding list structure. The approach described here is somewhat different in that as we first create a large data base (i.e., a node dictionary) for the representation of individual nodes and their feasible connections. A network is then represented not by a separate list structure but, symbolically, by a code which specifies the nodes and connections which define the network.

The basic reason for taking such an approach is that it allows a simple batched input of network data from which the user can select and modify pieces of the network which interest him. It is considered that the user will be concerned only with a portion of the total transportation network at any one time. For example, he might be interested in only one mode of transportation or he might be interested in only connections near a central location or between two specified nodes: The proposed method of representing the network facilitates the generation of such subnetworks.

Input Operation Data Structure

Input Data There are two distinct parts to the input operation: the first is concerned with the details of how information is taken from an actual map and transferred to an input format; the second is concerned with an input algorithm which takes the card input and constructs the required data structure. For any input map we assume that an overlay is available from which relative coordinates can be extracted. In other words, something like a variable size 1000 x 1000 grid is available. For each input node the following information is required:

1) NAME
2) COORDINATES
3) NUMBER OF ADJOINING NODES (i.e., the degree)
4) NAMES OF ADJOINING NODES
5) INDICATION OF IMPORTANCE
6) TYPE OF NODE

Only 5) and 6) need a further explanation. Certain nodes will be designated as important nodes and will comprise the entry nodes (i.e., the entry points for the user). When the user starts the system the first command is DISPLAY ENTRY POINTS. At this command all those nodes which have been coded as important will be displayed on the display system. During the use of the system the user can modify the entry list.

In general, there will be more than one type of structure contained in the input data map. For example, the map might be of an urban transportation system with more than one mode of transport or a city street map with a number of levels of detail. The coding for the type of node will contain the information as to what substructures a particular node belongs to. A possible format for a card input representing a node is given below.

8	10	24	30	32	40	48	56	64	72	80
NAME	DEGREE	COORD	TYPE	IMP	NAME	NAME	NAME	NAME	NAME	NAME

In columns 1-8 there is the name of the node. In columns 9-10 there is a digit from 1-6 giving the number of connecting nodes. In columns 11-24 there are the coordinates in format xxx.x,xxx.x. Assuming that at most 3 different types of structures are allowed, columns 25-30 would tell to what structures a node belonged to, the possible codings would be:

1
2
3
1,2
1,3
2,3
1,2,3

Columns 31-32 would contain a one if it is an entry node, a zero otherwise. The names of the connecting nodes would be given in blocks of 8 columns from columns 32 to 80. In this format six connecting nodes at most can be accommodated. The details of the input algorithm used for constructing the required data structure from the basic card input will not be presented here.

Data Structure The data structure for a node N will consist of a fixed length field and a variable length field. The fined length field will contain the following subfields:

i) N - name of node,
ii) D - degree of node (number of connecting nodes),
iii) C - coordinates for scope presentation,
iv) IND - indicator of whether node is in display list,
v) M - code for node membership in subsets of nodes,
vi) CONN - code for indicating connections that are being utilized.

The variable length field will have a subfield ADD for each connecting node containing the address of the field representing each connecting node. The entire field for a node having 3 connecting nodes is given below.

N	D	C	IND	M	CONN	ADD	ADD	ADD	

Language Development

We now must develop a language or notation for addressing and organizing sets of subfields of the data structure. The totality of representation for the nodes will henceforth be referred to as the node dictionary. Capabilities must be developed for:

i) modification of the dictionary,
ii) construction and representation of subnetworks for display and for input to the applications programs, and
iii) manipulation or modification of a given subnetwork.

Basic Notation The first step is to develop a notation for addressing the subfields of a particular node. Let us remember that the data structure for a node is:

NAME	D	C	M	IND	CONN	ADD_ _ _ADD

↑
X

Let X be the (leading) address of the nodal representation, then we define the following operators:

1) N (X) ; the name of the node whose address is X,
2) A (N(X)); the address of the node whose name is N (X) hence A (N(X)) = X and N (A(Y)) = Y where Y is the name of the node whose address is A (Y),
3) D (X) ; the degree of the node whose address is X, D is an integer giving the number of connecting nodes to N (X) and hence the number of address subfields in the variable length field, and
4) i (X) ; the address of the i^{th} node reachable from the node whose address is X.

The contents and addressibility of the remaining subfields will be discussed later. They are relevant to the construction of subnetworks from the dictionary.

Linkage Notation In the previous section we defined a notation for referencing the attributes of a single node. We will now expand the notation to represent sets of nodes and their connections. Previously we had i (X) where i is an integer from 1 to D (X) as the address of the i^{th} node reachable from the node whose address is X. We now define the natural extension.

1) ijk (X) = i (j (k(X)))
   as the address of the node which is i^{th} reachable from the node which is j^{th} reachable which is the k^{th} reachable from the node whose address is X. The generalized notation is $\underline{S}$ (X) where $\underline{S}$ is a vector of integers with the constraint that if $\underline{S} = S_n S_{n-1} \ldots S_1$ then $1 \le S_j \le D(S_{j-1} \ldots S_1(x))$ for j = 2, n. This condition says that for any given node no reference can be made to more nodes than are connected to it. A path is now easily represented by P ($\underline{S}$ (X)) where the initial node is N (X) and the final node is N ($\underline{S}$ (X)). To identify the first occurrences of distinct nodes in a path we define the logical vector.
2) $\underline{L}$ (P ($\underline{S}$ (X)))
   such that $L_i = 1$ if $S_i S_{i-1} \ldots S_1(x) \neq S_j S_{j-1} \ldots S_1(x)$ for any $j \le S_{i-1}$
   In other words $L_i = 1$ when $S_i S_{i-1} \ldots S_1(x)$ is the first occurrence of the node $N(S_i S_{i-1} \ldots S_1(x))$ in the path P ($\underline{S}$ (X)). From P ($\underline{S}$ (X)) and $\underline{L}$ (P ($\underline{S}$ (X))) it is easy to define.
3) SET (P ($\underline{S}$ (X))),
   the set of all distinct nodes in the path P ($\underline{S}$ (X)). Finally we define
4) T (X^V) as the set of distinct nodes at most v connections away from N (X). The construction of T (X^V) is not a simple matter and will be discussed in the section on algorithms.

Example of Notation In this section we present a simple network and give the dictionary for all the nodes and connections. The notation is then demonstrated by constructing the representation of subnetworks. Consider the network of Figure 10 below.

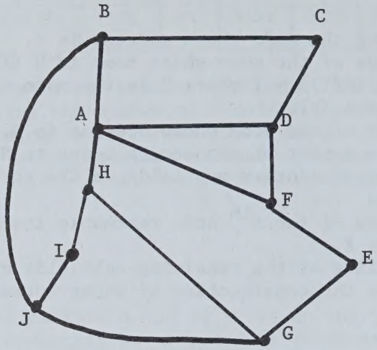


Figure 10

Remembering the data structure for a particular node is

NAME	DEGREE	CO-ORD	IND	M	CONN	N	N --- N

as defined in 2.2.3 a dictionary would be the following

ADD										
20	A	4					40	80	120	160
40	B	3					20	60	200	
60	C	2					40	80		
80	D	3					60	20	120	
100	E	2					120	140		
120	F	3					100	80	20	
140	G	2					100	160	200	
160	H	3					20	140	180	
180	I	2					160	200		
200	J	1					180	140	40	

The addresses could be almost anywhere as they need not be dense, sequential, or increasing. The particular ones chosen are only for aesthetic presentation.

Only the NAME, DEGREE and N fields are necessary when the full power of the notation is used to describe subnetworks. The other fields are necessary for the operational algorithms since only a part of the notation can be implemented on the machine. In other words a human intelligence is needed in going from a pictorial representation to a algebraic one when all the notation is utilized. The details of the actual algorithms are given in section 2.4. The following illustrate the notation:

N (60) = C
N (180) = I
A (D) = 80
A (F) = 120
N (A(F)) = F
A (N(60))= C

2 (40) = 60
3 (140) = 200
12 (20) = 1 (2(20)) = 1 (80) = 60

211 (60) = 80

12 (40) = 40
12 (A(B)) = A(B) = 40

The path from A to B to J to G is represented by P (231 (20)). A network can in general be represented by a set of paths. For example consider the subnetwork given in Figure 11.

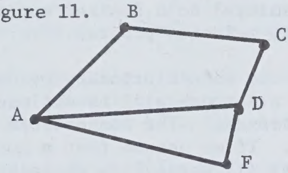


Figure 11

Two possible representations of the network are:

i) 232221 (20) - [ABCDAFD]
ii) 221 (20); 322 (120) - [ABCD]; [FADF]

An actual study of the use of such notations in the development of an algebra of network manipulation has not been attempted. It might very well be a fruitful area of research.

Operations and Algorithms

In this section we first give an overview of how networks are constructed from the basic data structure. Then a listing of the particular algorithms is given. A detailed specification of the algorithms will not be given because of space requirement.

Network Building In this section an outline of the procedure for the construction of subnetworks will be given. We start with the dictionary of nodes and an entry list containing the addresses of certain specified nodes. Originally the node in the entry list are specified at the time of input but the user has the option to modify the input listing. The purpose of the entry list is to give the user a set of starting points from which he can build any network. A subnetwork is built by first defining the basic set of nodes comprising the network and then adding the desired connections. The address of the nodes to be displayed (comprising the subnetwork) at any stage are contained in a sequential list represented by the vector U. The set of node addresses used as entry points are also in a sequential list represented by the vector ENT. The construction begins with the user command DISPLAY ENTRY POINTS which initializes U̲ to E̲N̲T̲.

At any stage in the construction of the network the nodes whose addresses appear in U̲ are displayed to the user. To modify the status of U̲ or the set of displayed nodes the user identifies a particular node by pointing to it. He then has the following command options:

1) REMOVE NODE,
2) REMOVE (ADD) ALL NODES WITHIN (n), and
3) REMOVE (ADD) NODE (S̲)

where n is an integer and S̲ is a vector of integers.

Option (1) removes the given node from the display list. Option (2) removes (adds) all those nodes that can be reached from the given node in at least (n) steps. Option (3) removes (adds) that node referenced by S̲ (X) where X is the address of the given node.

Once a basic set of nodes is decided upon the command DISPLAY CONNECTIONS is given, this command displays all the possible connections between the selected nodes. The user can then remove any of the existing links from the display or actually create new links or even new nodes. The last two operations entail changes to the node dictionary. The building of the network continues until the user is satisfied with the configuration. The displayed network is then transmitted to a desired applications program in a format suitable for use by the selected program.

Subnetworks are generated from the node dictionary by use of simple reference techniques. The nodes of a network will be defined by a vector U̲ which will contain the node addresses. The connections for each node will be specified by the CONN field. If we assume that a node will have no more than six connecting nodes. CONN (X) where X is the address of a node will be a digit from 1 to 63 representing what connections are included. For example if CONN (X) = 13 node N (X) is connected to N (1(X)), N (3(X), and N (4(X)). In general, for a node in a network we make reference to the individual binary components of CONN with the following notation:

In other words if N (U_i) is connected to the 2nd and 4th nodes in its address field

CON (U_i , 1) = 0
CON (U_i , 2) = 1
CON (U_i , 3) = 0
CON (U_i , 4) = 1, and
CONN (U_i) = 10

Hence a network can be described by 2 vectors, the vector U̲ of address and a connection vector C̲N̲ where CN_i = CONN (U). This simple reference code permits the easy storage of many networks referring to the same node dictionary. If many subnetworks are processed all referring to the same node dictionary this provides an easy way to keep a record of all the structures.

The algorithms of the system are defined to operate on the two vector quantities U̲ and C̲N̲. Construction and manipulation of network structures consist of monifying the two vectors.

Basic Operational Algorithms In this section we will consider the basic algorithms that operate on the two vectors U̲ and C̲N̲ representing a network structure. The first entries into U̲ are from the E̲N̲T̲ vector which is the address of the nodes specified as important. The manipulation algorithms can be divided into four classes - those which deal with the building of the set of nodes, those which deal with the establishment of connections between them, those which deal with modification of the node dictionary and those which deal with format conversion. The algorithms concerned with the building of the set of nodes are:

i) ALG I - adding those nodes to the list vector U which are within (n) connections from a given node
ii) ALG II - removing those nodes from the list vector U which are within (n) connections from a given node
iii) ALG III - adding a particular node to the list vector that can be reached by a specified path from a given node
iv) ALG IV - deleting a particular node from the list vector that can be reached by a specified path from a given node

The initial connections are specified by the field CONN corresponding to each node in the vector list U. The following algorithms are used for modifying the initial connections.

v) ALG V - adding a specified link from a given node
vi) ALG VI - removing a specified link from a given node
vii) ALG VII - adding a specified path starting from a given node
viii) ALG VIII- deleting a specified path starting from a given node

Added flexibility is given the system by allowing modifications of the node dictionary during the construction of a network structure for display purposes. The algorithms used for dictionary modifications are:

i) ALG IX - the creation of a new node
ii) ALG X - the removal of a node from the dictionary
iii) ALG XI - the creation of a link between the two nodes

The detailed operations of the display software system is not considered in this paper. Our analysis is concerned with preparing a network data structure for interfacing with a display system. A vector of coordinates will be the format when just the nodes are displayed while a set of coordinate pairs will be necessary when connections are to be displayed. The one basic algorithm needed here is to convert from the two vectors U and CN to the set of coordinate pairs. This will be ALG XII.

User Operations

In this section we list the actual commands available to the user of the system. The commands are as follows:

```
DISPLAY ENTRY POINTS
DISPLAY NODES
IDENTIFY NODE (point to node)
  REMOVE   NODE
  REMOVE   NODES (n)
  ADD      NODES (n)
  ADD      NODE  (s)
  REMOVE   NODE  (s)
  DESTROY NODE
  PUT IN ENTRY LIST
  REMOVE FROM ENTRY LIST
DISPLAY CONNECTIONS
IDENTIFY NODE
  REMOVE LINK (i)
  ADD LINK (i)
  ADD PATH (s)
  REMOVE PATH (s)
```

```
CREATE LINK TO (NAME)
CREATE LINK TO (point to node)
CREATE NODE (NAME)
COOR ARE (point to location)
M is ( )
```

Summary and Recommendations

A methodology has been developed which takes raw data about a transportation network and converts it into a data structure suitable for comprising the basis of an interactive network building system. The four parts of the analysis are:

1) input and data structure,
2) operations on the data structure as defined by user commands,
3) data interface with display system, and
4) data interface with applications programs.

The heart of the network manipulation system is the data structure and the operations defined thereon. Since more network data will be present than will be used at one time a new method of representation has been developed. A relatively stable "dictionary" of information is constructed and then networks are represented by a flexible code which defines subnetworks. Most modifications of the displayed networks are then performed on the code rather than on the basic data structure or dictionary. User commands are defined by algorithms which operate mainly on the code.

OPTIMIZATION OF PROGRESSION TIMINGS ON A UNIFORM GRID STREET NETWORK

JOHN B. KREER
Department of Electrical Engineering and Systems Science
Michigan State University
East Lansing, Michigan 48823

ABSTRACT

A computer control algorithm is described which minimizes a cost function that is a measure of total vehicle driving time in the network without requiring drivers to travel closer together in the progression band than is natural. It is shown that the timing constraints imposed by the grid configuration limit the realization of progression timing to (1) one critical street in each direction or (2) progression on all streets in three directions.

INTRODUCTION

Increasing population in urban areas, limited space for additional highway facilities, the escalating cost of highway construction and the desire to provide improved service to urban highway users has led a number of cities to install digital process control computers to coordinate their traffic signals. In such systems a number of electronic vehicle detectors are located on the principal streets under control to provide a closed loop control system. Most of the control algorithms used in these systems have been empirical in respect to the way traffic measurements were related to the selection of traffic signal timing patterns. The results of early experiments with such systems have shown that small changes in traffic signal timing can result in large improvements in traffic flow.

Experience has shown that better street utilization and user serivce will be benefited by setting the relative timing of traffic signals so that progression occurs which enables a driver theoretically to drive along a street and make a maximum of one stop for a traffic signal. However, there are several constraints on the selection of traffic signal timing for progression which limit what can be accomplished. In particular it will be shown that closed paths along which progression is to be maintained also cause some serious constraints on what can be accomplished. Since many central business districts where computer traffic control systems are being installed consist of a grid of intersecting streets, it is important to have an algorithm for selecting the optimum progression timing for a grid street network.

BASIC PROGRESSION TIMING CONCEPTS

In specifying a progression timing, three basic parameters are used--cycle length, split and offset. Cycle length is the time it takes for an individual traffic signal to progress through its complete sequence of phases. It is fairly obvious that for progression timing to exist for successive platoons a periodicity in the timing patterns is necessary which requires that all signals be operated with the same cycle length or cycle lengths that are integral multiples of the shortest cycle length. For most situations a single cycle length is employed for all traffic signals in an area. Split is the fraction of the cycle length in which the principal direction receives a green phase. Offset is the relative time displacement of the middle of the green phase at a given intersection relative to the middle of the green phase at a selected reference intersection. Specification of all three of these parameters defines the state of all traffic signals in the network.

A common measure of usefulness of a given progression timing is the bandwidth. Bandwidth is that portion of a cycle during which a car could start at one end of the street and, by traveling at the progression speed, go to the other end without stopping.

OPTIMIZATION CONSTRAINTS

In considering the optimum progression for a single two way street in reference 1 it was shown that in order for vehicles to be able to remain within the bandwidth without driving faster than is natural for the existing vehicle, spacing the progression speed and bandwidth must be chosen so that

$$y_d \leq 4\,B\,y_m \left(1 - \frac{v_p}{v_f}\right)\left(\frac{v_p}{v_f}\right) \qquad (1)$$

where

y_d is the demand flow rate for the street
B is the progression bandwidth in fraction of a cycle
y_m is the capacity of the street
v_p is the progression speed
v_f is the free speed of the street.

This constraint is based on macroscopic traffic flow models which have been reasonably well validated by field measurements.

A special situation with properly designed progressions for one-way streets is that the bandwidth is equal to the minimum green time.

If uniform splits are used along a one-way street, then the bandwidth and the split will be equal. In a one-way grid it can be seen that if a given split is justified at one intersection in terms of the demand on a given street, it will also be justified at the next intersection. Thus one can select only a split or bandwidth for the East-West streets and the remainder of the cycle becomes the bandwidth for the North-South streets.

A convenient approximate measure of total travel time in the network is given by

$$F = \sum_i \frac{y_{di}}{v_{pi}} \qquad (2)$$

where

y_{di} is the demand on the i th street
v_{pi} is the progression speed on the i th street.

The number of users is directly proportional to the demand flow rate and the average travel time per user is directly proportional to the reciprocal of progression speed provided that (1) is satisfied so that progression flow does in fact take place.

If the demands on all streets in a given direction are uniform, then the progression speed is chosen such that (1) is satisfied as an equality. Substituting the right hand side of (1) into (2) results in

$$F = \frac{4y_m}{v_f^2} \left[B_{EW}(2v_f - V_E - W_W) + B_{NS}(2v_f - v_N - v_S) \right] \qquad (3)$$

where the E, W, S, N subscripts refer to the bandwidth and progression speeds which apply to the four directions.

Since it is necessary that

$$B_{EW} + B_{NS} = 1 \qquad (4)$$

one can apply the method of Lagrange multipliers to minimize F subject to this constraint. This requires that

$$G = \frac{4y_m}{v_f^2} \left[B_{EW}(2v_f - v_E - v_W) + B_{NS}(2v_f - v_N - v_S) \right]$$

$$+ \mu \left[B_{EW} + B_{NS} - 1 \right] \qquad (5)$$

be minimized where μ is a Lagrange multiplier. When this minimization is done the result is that the bandwidths must be selected so that

$$\frac{B_{EW}}{B_{NS}} = \frac{y_{dE} + y_{dW}}{y_{dN} + y_{dS}} \quad . \tag{6}$$

This result is consistent with the usual practice of selecting splits in proportion to the estimated relative demands in the competing intersection directions.

Another constraint which must be satisfied by a progression timing for a grid network is making the sum of the offsets around any closed path equal to an integral number of cycles.[2] As one can see in what follows, this constraint generally proves to be very difficult to satisfy.

SELECTION OF THE OPTIMAL PROGRESSION

For a grid of one way streets the offset constraint necessitates the choice of one of two strategies depending on the relative distribution of demands within the network. Either one can obtain a progression timing on the heaviest demand street in each direction and apply some substrategy for the remaining streets or one can obtain progression for all streets running in the same directions as the three directions that have the heaviest individual street demands.

If the first strategy is selected, the splits would initially be selected using (6). The progression speeds for the four heaviest directions would be selected so that they are as large as possible but still satisfy (1). The cycle length would then be determined so that the transit time around the closed path at the progression speeds is some integral number of cycle lengths. The practical bounds on the minimum and maximum cycle lengths will limit the possibilities to a small number. With a one-way system of streets, the offsets can easily be determined once the splits, cycle length and progression speeds have been determined. The progression parameters for streets other than the critical four streets in most cities will be dictated by peculiar features in the local geography. In other cases the remaining streets could be treated as subnetworks bounded by the critical streets with the same strategies employed on the subnetworks.

If the second strategy is selected, progression would be established for the three directions corresponding to the three streets having the heaviest demands. The splits will again be determined by (6). The streets for one direction will be orthogonal to the other two. The offsets required for progression on the streets in this direction will cancel out in the sum of the offsets around closed paths in the network. This means that the sum of the offsets in the other two directions must equal an integral number of cycle lengths. This constraint can then be used to determine the cycle length. Once this has been determined it is routine to determine the offsets so that the progression speeds on the three highest demand streets just satisfy the equality in (1). The progression speeds for other streets in the same directions will be somewhat lower than the demand on these streets require if these streets could be decoupled from the grid network. However, the discrepancy will be small unless the demands are much lighter. In this latter case the number of cars being penalized is small and an overall network optimum or near optimum should result as far as minimizing total travel time in the network.

The choice of which strategy is to be employed for a given demand distribution in the network is based on the calculation of a measure of total vehicle travel time

$$T = \sum_i \frac{d_i}{v_i}$$

where

d_i = the measured demand on i th street

v_i = the progression speed on the i th street if the strategy allows progression otherwise 10 miles per hour or whatever nominal speed local measurements indicate for a non-progression street.

The strategy which results in the lowest value of T would then be implemented by the control computer for the timing of the network signals.

EFFECTS OF TWO WAY STREETS

The above strategies have been described assuming that the grid consisted of only one way streets. If there are two way streets in the network, then the split and bandwidth are no longer equal, and some difficulties can be encountered in satisfying the constraint on the offsets around closed paths. It has been shown[3] that on a two way street the total bandwidth that can be achieved is a constant and that within limits the bandwidth can be traded linearly between the two directions. For most street geometries this total bandwidth is equal to the split. If only a few two way streets exist in the grid, it may be possible to select the timing parameters for these streets by the method of reference 1 and then proceed to determine the remaining progression timing parameters as described above.

EFFECTS OF NONUNIFORM STREET SPACING

The effects of nonuniform street spacing will prohibit the use of the first strategy since the distance around various closed paths in the network will not be integrally related and thus a suitable cycle length cannot be selected for all closed paths. In the second strategy progression is maintained around only one closed path so that no difficulty should be encountered in using it in a nonuniform street grid.

CONCLUSIONS

It has been shown that the best that can be obtained in attempting to maintain progression on a grid of streets is to maintain progression on the most heavily traveled street in each direction or to maintain progression on all streets in the three most heavily traveled directions. With computer control of traffic signal timing it can be seen that much more responsive modes of traffic control can be obtained.

ACKNOWLEDGEMENT

The research described in this paper was supported by the Division of Engineering Research at Michigan State University, East Lansing, Michigan.

REFERENCES

1. J. B. Kreer and W. D. Panyan, "Optimal Selection of a Progression for a Two-way Arterial Street," Highway Research Record No. 286, 75-81, 1969.

2. W. D. Panyan, "Urban Traffic System Simulation and Control," Ph. D. Thesis, Michigan State University, 1969.

3. J. D. C. Little, B. V. Martin, and J. T. Morgan, "Synchronizing Traffic Signals for Maximal Bandwidths," Highway Research Record No. 118, 21-47, 1966.

RESERVATION SYSTEMS FOR SYNCHRONOUS TYPE GUIDEWAY AUTOMATION*

LOUIS T. KLAUDER, JR.
Louis T. Klauder and Associates
Philadelphia, Pa.

ABSTRACT

Suppose that the vehicles on a highway network are remotely controlled and move in synchronism. There must then be some logical procedure for deciding when to add successive vehicles to the flow. Two such procedures have been described previously. The present work attempts to distinguish all the procedures likely to be used for this purpose.

INTRODUCTION

Current transportation practice presents some problems which guideway automation may help to aleviate.

In present day American experience highway lane capacity is found to have a limit whose dependence on speed is roughly as follows:

Speed (m.p.h.)	Capacity (cars per lane per hour)
35	2200
50	1400
60	500

On the other hand, it is expected that with a traffic speed of 60 mph the capacity of an automated lane could be in the neighborhood of 8,000 vehicles/hour. In a congested urban area such an improvement would be quite beneficial both in reducing travel time and in reducing the amount of land needed for roadways. A reduction in accidents would be expected also.

As public transit seeks to compete with travel by private auto, it is faced with the need to provide frequent service at a very large number of points. Provision of such service would require use of a large number of small vehicles. The economics of such an operation would presumably be less difficult with automation than if there were an operator on board each vehicle. It is partly with this in mind that the U.S.D.O.T. is sponsoring the demonstration of an automated personal rapid transit system in Morgantown, W. Va.

There are several principles which can be used for providing automated longitudinal vehicle control. Automation could be provided using railroad type signaling which would permit vehicles to travel at assigned speeds so long as they did not approach preceeding vehicles too closely. Greater lane capacities could be realized if the fixed block type signaling used on railroads were replaced by systems which sensed speed differences as well as spatial separations between successive vehicles. This type of operation is referred to as automated car following and has received a fair amount of attention recently.[1] The present paper considers an alternative control procedure in which each vehicle is made to follow a space-time trajectory to be planned out before the trip begins. Systems of this type are sometimes referred to as point follower systems. It seems intuitively clear that the most efficient way to arrange for point following is to divide each guideway lane into conceptual slots which move with the traffic and each of which can be either occupied by a vehicle or else empty. The vehicle in an occupied slot is then controlled so as to stay in its slot as the latter moves along. The slot passage time is made to be constant throughout the system, and an automated guideway of this type may therefore be referred to as a synchronous slot guideway (SSG).

---

*This work was done while the author was at the General Motors Research Laboratories, Warren, Michigan.

Assuming a guideway network and vehicle controls to be available, there is need for a reservation system, i.e., a set of logical procedures for deciding which slots each new trip should use. Since new trip requests will be somewhat random, it is quite unlikely that the slots of an SSG could ever be all filled at once. How efficiently the guideway can be utilized will depend on the specific reservation system used for scheduling the trips.

Contributions to the SSG concept have been made by various authors.[2] However, the first reasonably complete description of a reservation system was that developed by the T.R.W. Systems Group.[3] Another system has since been described by the Alden Corp.[4]

The present talk is a partial report of work[5] which sought to discern all the procedures likely to be useful for scheduling trips on an SSG. Some new procedures have been found. However, no attempt has been made to develop quantitative information on the relative effectiveness of the various procedures currently available.

DESCRIPTION OF GUIDEWAY

The guideway will be referred to as consisting of sections joined to one another at intersections. Each section can have various on and off ramps. Each section into which a given section can empty will be labelled as one of the exits from the given section. The term trip will refer to the use of a specific slot on each of the sections to be traversed.

FUNCTIONS OF RESERVATION SYSTEM

The principal functions of the reservation system will be as follows:

1. Selection of a route through the network for each new trip.
2. Keeping track of whether or not each slot is available for use by a new trip.
3. Keeping track for each slot which is occupied or reserved of where the vehicle is to leave it.
4. Fitting new trips into those combinations of open slots which are available.
5. Providing for equitable dispensing of the capacity of each section and off ramp of the network so that vehicles entering the network at one point do not have an undue advantage over those entering at other points.

Attention will be directed here to items 2, 3, and 4. The problem of choosing routes through a network will not be considered because the techniques available for route selection are well described in the literature. Capacity allocation is very important in practice, but will not be discussed here because we have not found any attractive alternative to the capacity allocation procedure set forth in the TRW report.[1] The second function will be discussed after the third one as this order permits the discussion to be somewhat simpler.

STORING EXIT LABELS

The simplest way to describe a given route is to enumerate the sequence of exists or turns which occur along it. The last exist in any turn sequence will be via an off ramp, and any other turns will be exists from one section of guideway to another.

During normal operation the turn sequence of a given trip is used only for signaling the turns which are to occur as the trip progresses. For this use it would be sufficient to have the whole list associated simply with the current location of the vehicle.

However, if for any reason a section or off ramp becomes unusable, it will be desirable for the control computer to be able to try to reschedule all the trips which are effected. In order to do so, the computer must be able to reconstruct each trip which was scheduled to use a disabled off ramp or section. The Alden Report[4] describes one arrangement for doing this. The arrangement to be described here is equivalent in its results but is expected to be somewhat more efficient in practice.

For each section the reservation system is to maintain a one-dimensional array having one element associated with each slot on the section and having additional elements for those slots which will be appearing on the section in the near future. For each slot which is occupied, the corresponding element is to contain the label of the turn by which the trip in question will leave that slot. These arrays will be referred to as turn loops and their elements will be referred to as turn slots. These arrays will permit reconstruction of a trip using any given slot and will also serve to trigger all the turns.

As each new space-time slot enters its section, the correspondence between the turn loop and its section can be preserved by incrementing an index identifying the turn slot associated with the space-time slot currently at the beginning of the section. This saves having to move exit labels around on the turn loops.

BASIC RESERVATION ARRANGEMENTS

We adopt the following convention. Namely, reservations for a new trip are not to be made until after the first slot to be used by the trip has entered its section. One benefit of this convention is that it permits the turn loops to be kept as short as possible.

The basic reservation arrangements are as follows:

Simple Arrangement

The idea of this arrangement is to avoid having to keep track of more than one reservation at a time for any given slot. Since it would be undesirable to have one reservation of a slot preclude its use by an earlier trip which would vacate the slot prior to its use by the trip in question, the following rule applies. Reservations for a new trip are not to be made until the first slot of that trip has passed the last on ramp above the last off ramp above the on ramp at which the new trip is to begin.

Turn labels are to be inserted in turn slots as stated above. However, when the reservation is made, the turn slot corresponding to the first slot of the trip may still be occupied by the last exit label of a preceeding trip. Thus, the first exit label of the new trip is stored in a separate "load list" associated with the new trip's on ramp. The new trip's first exit label is then transferred to the proper turn slot when the vehicle is loaded. (The load list serves the additional purpose of triggering the loading.)

In addition, an extra "reservation bit" is added to each turn slot. This bit is changed from 0 to 1 when a new reservation is made and is changed back from 1 to zero when the vehicle which it represents passes the last on ramp above the exit at which the trip will leave the section. It is to the reservation bit that the reservation system makes reference when it wants to determine whether or not a given slot is free.

Control Point Arrangement

There are two arrangements by which a reservation system can keep track of more than one upcoming use of a given slot. One is the arrangement presented in the TRW report.[3] Since the other arrangement[5] maintains equivalent information and produces identical results, the TRW arrangement is the only one which will be described here.

The idea is to identify all the points in the network at which conflict could occur and then to keep track for each such point of whether or not each slot to pass that point will be occupied when it does pass. The points in question will be referred to as control points, and the arrangement will be referred to as the control point arrangement.

A control point is established at each on ramp and at each merge (where two or more sections flow together to form another section). In order to check on whether or not a given combination of slots can accomodate a new trip, the reservation system must check slot status at each of the control points which the trip will pass. For this purpose the computer is to maintain a list of reservation bits for each control point. Each reservation bit of a given list represents one space-time slot which will pass that control point. To make reservations for a new trip the system changes the proper bit from 0 to 1 for each control point which the trip will pass.

As before, the exit labels of a new trip are placed in the turn loops except that the first exit label is kept in a load list at the new trip's on ramp until the vehicle is actually loaded.

This arrangement needs a couple of refinements as follows. First,[3] if there were two control points with no off ramp or diverge between them, then a new trip could use a given slot passing both of them if and only if the slot were free at the lower of the two points. Therefore the upper of two such points may always be discarded. Second, for sections of a new trip which are subsequent to the first one, the only reservation bit which needs to be set = 1 when the reservation is made is the first one on the section. Then, each time that the vehicle is about to enter a new section, the remaining reservation bits for that section can be set on the basis of the exit label which tells where the vehicle will leave that section. This refinement affords a substantial reduction in the number of reservation bits which need to be maintained.

Consider a trip which will pass over 6 sections on which the reservation densities are currently

0.9, 0.6, 0.8, 0.7, 0.5, and 0.3.

The probability that any 6 properly aligned slots will all be open is

$$(1-0.9)(1-0.6)(1-0.8)(1-0.7)(1-0.5)(1-0.3) = \frac{1}{1200}$$

Assuming 40 ft. slot length and 60 mph speed, the time required for 1200 slots to go by is about 9 min. Having to wait an average of 9 minutes for admission to the network would be quite unacceptable. What is needed are ways of fitting trips into combinations of slots which are not perfectly aligned. Thus, the system needs to be able to move a vehicle from one slot to another while its trip is underway. Assuming that the physical facilities are designed to permit this, there are four ways in which they can be utilized.

Inter-Shift Modification

Suppose that a vehicle can be shifted by 1 slot up or down stream each time that it passes from one section to another. Then in the preceeding numerical example, the probability that reservations can be completed beginning from a slot picked at random will be somewhat greater than

$$(1-0.9)(1-0.6^3)(1-0.8^3)(1-0.7^3)(1-0.5^3)(1-0.3^3) = \frac{1}{47}$$

so that the average waiting time is reduced to about 20 sec. This modification is obviously pretty effective. When this modification is used, it is necessary to record the slot shifts along with the exit labels. It is sufficient for this purpose simply to enlarge the turn slots so that each of them can store a slot shift indicator as well as an exit label.

Intra-Shift Modification

A vehicle could also be shifted from one slot to another while traversing a single section. However, this modification would be applicable only if the reservation system were keeping track of more than one reservation at a time per slot. Each location at which intra-shifting would be useful would lie between an off ramp and an on ramp with the latter being the lower of the two.

To maintain the bookkeeping, each intra-shift is recorded as an exit followed by a re-entrance. Thus, the first exit label is for the off ramp immediately above the shift. Its identity as a shift rather than an exit is signaled by the shift index attached to it. The label of the eventual exit is placed in the load list associated with the on ramp immediately below the shift point and is transferred into the proper turn slot when the shift takes place.

Rescheduling Modification

If while trying to make reservations for a new trip the reservation system encounters occupied slots which it cannot circumvent using the slot shifting available, it could consider altering one of the previously scheduled trips so as to make an opening for the new one. Simple procedures of this sort are not hard to devise. They might or might not prove to be beneficial.

Cycle Modification

The cycle modification was presented in the TRW report.[3] It is so named because it groups adjacent slots into cycles with typically 4 or 5 slots per cycle. It is similar to the inter-shift modification mentioned above in that it provides for slot shifting between sections. However, instead of worrying at reservation time about just which slots each trip is to use, the cycle modification just checks to be sure that the cycle will be able to hold the vehicles scheduled to be in it. Slot shifts are then decided upon and carried out just prior to merges so that collisions are avoided.

The list of reservation bits associated with each section (or with each control point in the case of the control point arrangement) is replaced by a list giving the scheduled occupanices of the cycles. The procedure for keeping track of the route of a trip is as follows: the label for the exit from the last section of the trip is placed in any empty turn slot associated with the trip's last cycle; in any free turn slot of the cycle to be used on the next to last section of the trip are placed both the label of the next to last exit and also an index identifying the slot into which the last exit label was placed; etc. The label of the first exit is kept in a load list until the trip begins, at which time it is placed in the slot corresponding to the vehicle's actual location. Whenever a new cycle enters the beginning of a section, the labels in the corresponding turn slots are rearranged so that each one is placed in the turn slot corresponding to the space time slot occupied by the vehicle which it represents.

Comparison

The cycle modification tends to combine all the features of the other three modifications, and it might therefore appear to be obviously superior. However the cycle modification has two features which may be disadvantageous. The first feature bears on the question of reliability; namely, some of the collision avoidance decisions are not made until the trips involved are already underway. The second feature has to do with scheduling efficiency; namely, the cycle modification incorporates the constraint that vehicles not be shifted across the boundaries between cycles.

SUMMARY

Taking the various procedures outlined above, the eight combinations which appear interesting are as follows.

Modification	Arrangement Simple	 Control Point
inter-shift	1	2
inter-shift plus intra-shift	*	3
inter-shift plus rescheduling	4	5
inter-shift plus intra-shift plus rescheduling	*	6
cycle	7	8

*intra-shift not relevant with simple arrangement

Finding out how these various alternatives would compare in practice will depend on detailed simulations.

REFERENCES

1. Cf. R. E. Fenton, IEEE Transactions, VT-19, 153-161 (1970).

2. Baumann, D. M., et al. in A.S.M.E. National Transportation Symposium, San Francisco, Calif., May, 1966, pp. 199-209; U.S. Dept. of Commerce documents as follows:
   PB# 178-283, pp. 78-80
   PB# 178-275, Case Study A: Network Cab.

3. "Study of Synchronous Longitudinal Guidance as Applied to Intercity Automated Highway Networks," T.R.W. Systems Group, Washington, D. C. (1969). (U.S. Dept. of Commerce document PB# 188-582)

4. "Alden Capsule Transit System Control Subsystem and Baseline Definition," Alden Corp., Westboro, Mass. (1970). (U.S. Dept. of Commerce document PB# 192-737, Chap. 3)

5. Klauder, L. T., Jr., to be published elsewhere.

A SYSTEM ANALYSIS OF UNIVERSITY TRANSPORT NEEDS

J. E. Gibson
School of Engineering
Oakland University
Rochester, Michigan

A. W. Turski[*]
School of Engineering
Oakland University
Rochester, Michigan

INTRODUCTION

The transport needs of a medium-size suburban university are analyzed and the resulting action plan is presented. Oakland is a state university located 25 miles north of downtown Detroit, and is not served by any public transport (see fig. 1). Of the more than 7000 students, faculty, and staff, approximately 5000 commute daily by private auto. With an annual growth rate of more than ten percent the university is faced with major dollar and esthetic costs for additional parking and continued implicit discrimination if transport policy is not changed. Students, faculty and staff living on and off campus are faced with the high cost of owning and operating personal automobiles.

This paper analyzes the origin/destination of trips in this many-to-one situation and then turns to an examination of a variety of possible transport mode solutions. Based on the consensus obtained in a series of public presentations to various decision groups, two prime candidate modes were selected for implementation in the action phase. University organized car pools will provide an immediate solution and Dial-A-Bus system will follow in January, 1972.

PRESENT TRANSPORT AT OAKLAND

Oakland University with over 2,800,000 trips per year to and from its campus is approaching a transportation crisis. Almost all of these trips are by personal automobile. The campus is not served by any public transportation. To the present, the university has followed a plan of building surface parking around the campus. Within five years all available land within 10 minutes walking distance of the center of campus will be used up and parking structures will have to be built. To amortize these structures, we calculate the parking fee would have to be increased from the present $16.00 to $75.00 per year. Presently an average of 1.2 people occupy each vehicle coming to campus. If the number of people per vehicle could be increased, traffic and parking demands could be reduced significantly. Presently Oakland is limited to students who can afford to own and/or operate cars. Thus many students who cannot afford cars find themselves stranded because of the lack of transportation to off-campus areas. One resident student said last spring that he had not been off campus for two months.

There exist promising transportation alternatives which could improve Oakland's transportation and make better utilization of parking areas.

---

[*] Presently with UniRoyal, Inc., Detroit, Michigan.

The possible application and implementation of these systems at Oakland University is the thrust of this paper.

A general planning goal for Oakland University is the following: to increase and equalize the educational experience of the people in the Southeastern Michigan Region by giving them an opportunity to utilize the resources of the university. Some of the resources are the academic program at Oakland, the Continuing Education program, the Sports and Recreation program, Meadow Brook Music Festival and Theater. The key that enables the people of the region to interact with and utilize the resources of the university thereby enabling the university to interact with and utilize the resources of the region is transportation.

Table 1 shows how many vehicle trips are generated by each resource of the university:

	Estimated Number of People Per Vehicle	Number of Vehicle Trips Per Year
Academic (Students, Faculty & Staff)	1.2	1,670,000
Continuing Education	1.0	210,000
Sports & Recreation	2.2	35,000
Meadow Brook Music Festival	2.2	116,000
Meadow Brook Theater	2.2	116,000
TOTAL		2,147,000

Table 1: Number of Vehicle Trips Per Year To and From Oakland

Figure 2 shows trip origins; each line represents about 100 people.

ANALYSIS OF FUTURE OPTIONS

The possibility of rapid rail transit serving the university in the near or medium range future is remote indeed. Therefore, we limit the analysis to transportation systems which operate on roads.

List of Alternatives

- Personal Automobile
- Shared auto or some form of car pool
- Dial-A-Ride or some form of a bus service on demand
- Express bus service picking people up at central points
- Fixed route, fixed schedule bus system
- Fixed schedule, flexible route
- Taxi
- Motorcycles
- Walking
- Other

How are trade-offs determined for the above systems? Who makes these trade-offs? The decision-maker is the user of the transportation system. Therefore, early in this system analysis riders were made aware of the costs and characteristics of the above systems by means of a three-day interactive transportation exhibit. Ford Motor Company permitted on campus a demonstration of their prototype Dial-A-Ride vehicle. Newspaper articles and pictures followed both in the school papers and the Detroit News. A special presentation was given to the President's administrative group.

Presently, there is one parking space provided for every faculty and staff member and one space for every 2.9 commuter students. Therefore, there are 1791 commuter spaces for 5200 commuters and 1155 faculty and staff spaces. Lots have been built on the west, north, and south portions of the developed campus. After an additional 2,000 spaces are built in the south end of campus, Oakland will have to build parking structures or bus people from outer lots in order to stay within the constraint of having parking located within 10 minutes of the center of campus. Using the estimated growth figures of additional 1200 students per year, Oakland will need to build parking structures in the Fall of 1976. We calculate the parking fee will have to be raised from $16 per year to $75 per year in order to amortize the structures. By the fall of 1977, if Oakland follows its present policy, $2,500,000 will have to be invested in parking lots and structures. Table 2 shows the amount of capital and maintenance expenditure for parking lots in the next seven years. The negative ecological effects of additional parking lots cannot be ignored.

	Total needed if present policy continued	Cost	Area in Acres
1971	256 (In planning 100 needed)	76,800	0.7
1972	336	100,800	3.2
1973	595	178,500	3.8
1974	597	179,100	3.8
1975	530	159,000	3.4
1976	601	901,500 - Parking structure	
1977	601	901,500	

Table 2: Parking Needs and Costs for the Next Seven Years

The cost of owning and operating an automobile for two semesters of travel to and from school is $500 to $1,000, depending upon the length of the trip. It costs at least 10¢ per mile to own and operate an automobile. Table 3 shows how these costs are broken down:

Gas	= 2¢ per mile
Oil, parts, and repairs	= 3¢ per mile
Insurance	= 2¢ per mile
Depreciation, license and title	= 3¢ per mile
	10¢ per mile total

Table 3: Costs of Operating an Automobile

A system analysis showed that the student, faculty or staff member could reduce his transportation costs by riding to school with other people and sharing the cost. The university would also save money because the need for parking lots would be reduced. The various ways of sharing vehicles was shown in the exhibit and presentations. A consensus was formed and the decision was made to implement two of the systems at Oakland.

TWO PRIME CANDIDATES

First, it was decided that a cooperative car pool system shall be organized at fall registration by Commuter Services personnel. With careful planning and publicity, the initial conventional car pool system will lead to a cooperative car pool system.

The cooperative car pool gives the student greater flexibility by enabling him to come to and go from campus in different cars. It usually operates with about 100 cooperating students and presents a tricky scheduling problem, but such a car pool has been working at Wayne State University in Detroit for about ten years.

Students tentatively are set up to go home in a certain car at a certain time, but every student has a list which tells them how many cars are going to his zone and at what time. This is the main advantage of such a cooperative car pool, because arrivals and departure times at school are scheduled by student for his convenience. The student is not tied to the rigid and perhaps inconvenient schedules of a conventional car pool, but he realizes all the cost advantages of the conventional car pool.

The Costs of a Cooperative Car Pool

If a student uses one of the family cars once a week to drive, as is the usual case in the Wayne State car pool arrangement, the student must pay only the gas for that day and the 50¢ parking fee at Wayne. Since it is about a 35 mile round trip from Grosse Pointe to Wayne, for example, gas costs about 70¢ per day, or a total cost of $1.20 per day, including parking. Since a student drives only one day a week, it only costs him $12.00 a quarter to get back and forth to school. If the student owns a car and only drives it once a week then fixed costs of depreciation and insurance of about $140.00 for a quarter must be added. The repairs and maintenance cost, 3¢ per mile, would be $10.50.

It is possible to be a "non-driver" member of the car pool. A non-driver pays about $40 a quarter to the Car Pool bank account. To balance out the non-drivers there are drivers who drive more than one day per week receiving $40 at the end of the quarter for the extra day he drove each week. The cars are usually fully loaded and it takes 20-30 minutes to pick up 3-5 students. The penalty for not picking up people is a $5 fine. If the fine is not paid the student is kicked out of the car pool. It is important to note if 10% of the students could be interested in some form of car pools, then 300-400 fewer parking spaces would be needed at a savings to the University of $90,000 to $120,000. Inducements such as special parking privileges for car pool drivers will be used to encourage cooperation.

The other system which will reduce the need for parking lots and provide the students, faculty, and staff with relatively low cost transportation service is Dial-A-Ride. Dial-A-Ride, which Oakland plans to implement on January 1972, is a flexibly routed bus service scheduled on demand. There are three basic steps in Dial-A-Ride Service:

(1) A request for service.

(2) Vehicle dispatched and routed to origin of user.

(3) Customer is picked up at his origin and dropped off at his destination.

The request for service can be made on a contract basis, or a day or even 30-60 minutes before the customer wishes to be picked up. Vehicle dispatching will be done manually. This is a many-to-one service because people come from many origins to one destination, the University. The Dial-A-Ride will be used by faculty, students and staff commuting to the university and by resident students who live on campus and want to get off campus. The service will probably use 8-15 passenger Ford Econoline type vehicles, with the headroom extended.

An important feature is the first class service in the passenger compartment, which has been designed for maximum passenger comfort. The seats are individual bucket seats with foam padding and high backs. The carpeted floor looks better than a rubber floor and helps deaden sound. Air conditioning will assure year-round comfortable interior temperature. Individual reading lights will be provided, thus enabling the passenger to study or read on his way to school. The area to the right of the driver provides a space to store packages while enroute.

The cost of Dial-A-Ride service from the Rochester or Pontiac area would be in the area of 50¢; from Birmingham or Troy - 75¢; from Royal Oak or Ferndale - $1.00; and from Detroit would cost $1.50 or so. The distance traveled and the number of passengers picked up, and level of service are some of the key factors in costing out such a system. Also, Dial-A-Ride will be able to operate as an express bus service to central pick-up points and as a regular fixed route as the demand for service develops.

Kent State University has shown that a bus system scheduled on demand can be successful. Their buses for the most part, operate on fixed schedules and fixed routes patterned according to demand. Ridership has increased to 22,000 riders per day. The system operates 25 buses logging a total of 1,800 miles per day. The service is almost completely run by students under the direction of a full time director. Thus, the students who manage the system are provided with an practical educational opportunity and students who work driving buses, etc., are provided with on campus employment. The student payroll is $160,000 or 40% of the collected revenue.

Oakland University will be an ideal place to carry on a controlled Dial-A-Ride experiment. First, students involved in the research will be compensated with university credit, thereby saving costly professional fees. Such a project will be an exciting educational experience for the student. Second, operating expenses are lower because of relatively low student labor costs. Third, the demand at a university is spread over the day, thereby making efficient use of equipment.

The Dial-A-Ride project at Oakland will be multidisciplinary in nature, involving faculty and both graduate and undergraduate students. People in management will work on the managing of the system. Psychologists and sociologists will study attitudinal changes in the implementation process. The social impact of the system will also be investigated. Engineers will work on the scheduling and routing, and the overall design of the system.

The cost of operating Dial-A-Ride buses is 15-20¢ per mile and driver costs are about $2.00 per hour, using student labor. Assuming eight hours of bus operation per day, five days per week, it will cost about $50 per day minimum to operate and pay the driver for each bus. The cost for two semesters for each bus including operating capital, and driver costs is $7,500. This figure does not include the cost of dispatching, management or research activities. To offer minimal service and serve one area well, it will take 2-3 buses. Two buses could be used if a motor pool could be depended upon when one vehicle was out of service. Table 4 shows a tentative budget for the pilot project. Funding for the project is now being sought in the public and private sector.

Manager	$10,000
Dispatcher	6,000
Three buses	23,000
Research	10,000
Miscellaneous & Contingency	5,000
Final Report	1,000
	$55,000

Table 4: Tentative budget for Dial-A-Ride with Minimal Service

One of the goals of the demonstration project will be to establish guidelines for implementation of Dial-A-Ride service at other universities and urban activity centers. During fall 1971, detailed design of the Dial-A-Ride experiment, including field test of the dispatching algorithm will be carried out.

REFERENCES

1. Golob, T.F., Canty, E.T., Gustafson, R.L., Vitt, J.E., An Analysis of Consumer Preferences for a Public Transportation System, GM Research Labs., Warren, Michigan, October 26, 1970.

2. Guenther, K.W. and Givens, W.E., "The Courier: A Prototype Vehicle for Dial-A-Ride Service" Society of Automotive Engineers, Paper No. 700186, New York, 1970.

3. Guenther, K.W., The Mansfield Dial-A-Ride Experiment, Ford Transportation Research and Planning Office - 70-7.

4. Guenther, K.W., Public Transportation Service Innovations: What Are The Lessons for Dial-A-Bus?, Ford Transportation Research and Planning Office - 69-18.

5. Johnson, Johnson, and Roy, Inc., Guidelines for Continuing Physical Development, Oakland University, Ann Arbor, Michigan, January 1971.

6. Kent State University, paper explaining their Transportation, Office of Parking and Traffic, courtesy of Mr. Michael Blurton, Associate Director, Kent State University, Center for Urban Regionalism, Kent, Ohio.

7. Oakland University Campus Guide.

8. Ward, R.F. and Kurz, T.E., "The Commuting Student", A study of facilities at Wayne State University, final report of the Commuter Centers Project supported by Educational Facilities Laboratories.

9. Gibson, J.E., "The City As A System", Oakland University, School of Engineering.

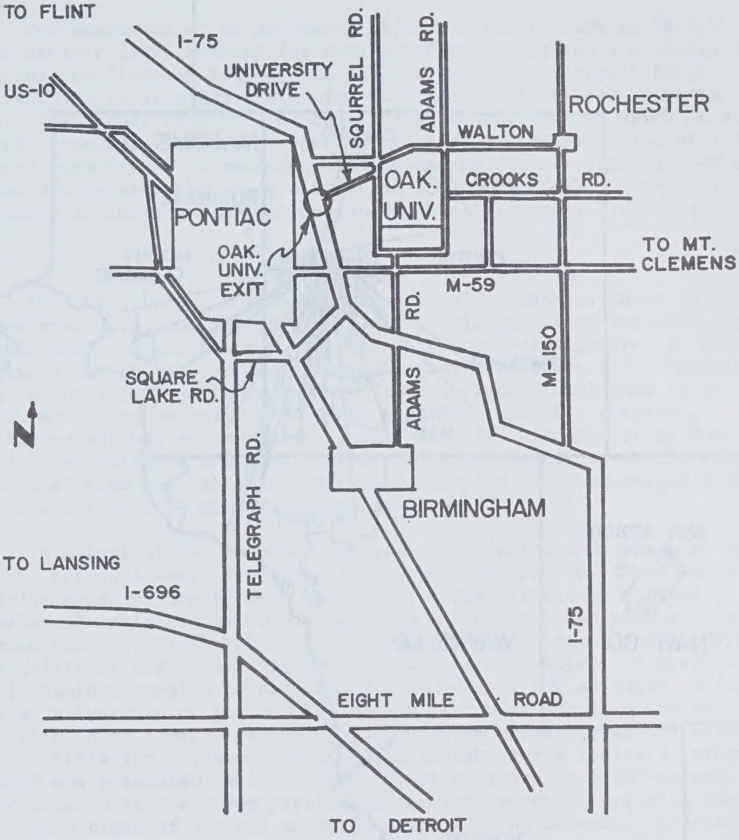

TO FLINT
I-75
US-10
UNIVERSITY DRIVE
SQURREL RD.
ADAMS RD.
ROCHESTER
WALTON
OAK. UNIV.
CROOKS
RD.
PONTIAC
OAK. UNIV. EXIT
TO MT. CLEMENS
M-59
RD.
M-150
SQUARE LAKE RD.
ADAMS
N
TELEGRAPH RD.
BIRMINGHAM
TO LANSING
I-696
I-75
EIGHT MILE
ROAD
TO DETROIT

FIG. 1

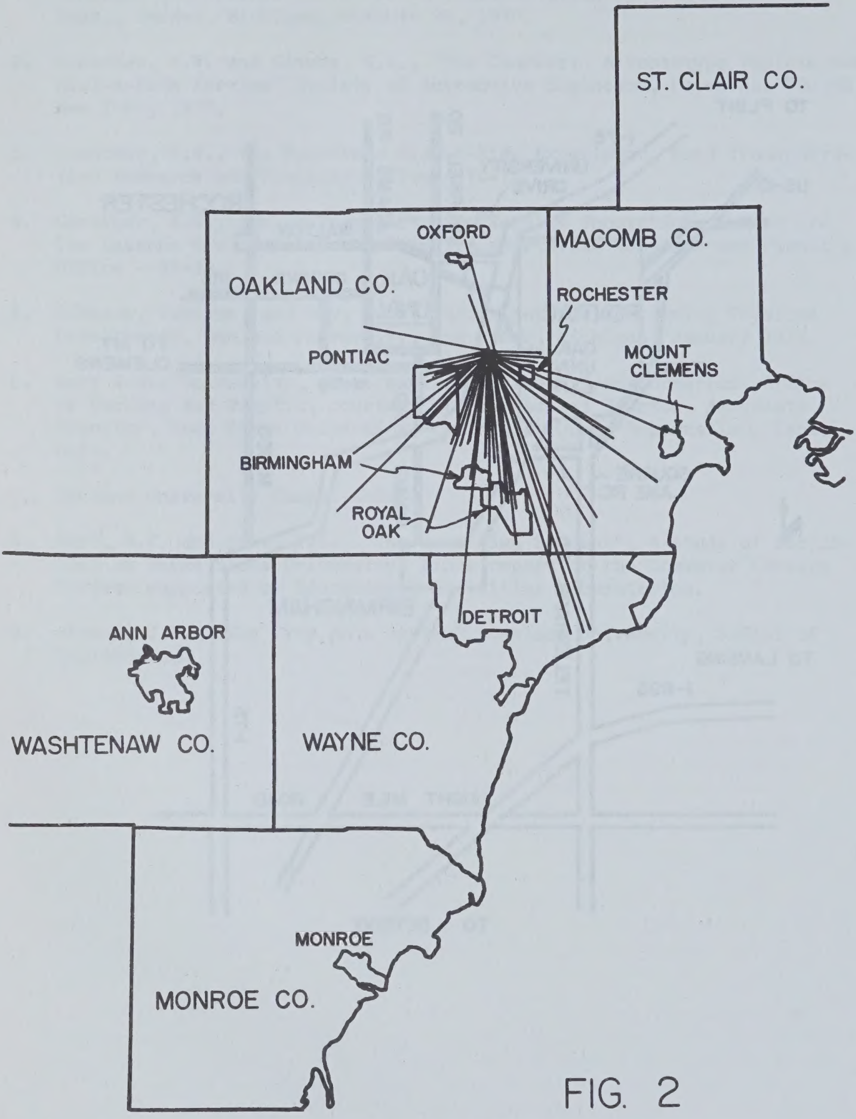

ST. CLAIR CO.
OAKLAND CO.
OXFORD
MACOMB CO.
ROCHESTER
PONTIAC
MOUNT CLEMENS
BIRMINGHAM
ROYAL OAK
ANN ARBOR
DETROIT
WASHTENAW CO.
WAYNE CO.
MONROE
MONROE CO.

FIG. 2

NETWORK EFFECTS ON THE INTERFACE PROBLEM FOR AN AUTOMATED HIGHWAY SYSTEM

ROBERT G. STEFANEK
Transportation Research and Planning Office
Ford Motor Company
Dearborn, Michigan

ABSTRACT

The operation of an automated highway network, such as in the Dual-Mode Vehicle (DMV) concept for such a network, requires a strategy for coordinating the flows of vehicles in the network so that smooth passage for all vehicles is attained while high utilization of network links is achieved. This paper examines the capacity of an entrance to a DMV network in a manner consistent with a proposed coordination strategy. The effects on entrance capacity of parameters which depend on vehicle capabilities and DMV system design are studied as a function of vehicle flows in the network. The influence of these parameters on network utilization is also treated.

INTRODUCTION

The DMV concept provides for operation of vehicles under automatic control on specially constructed "guideways", while the same vehicles could be operated under manual control, on the usual types of highways, in order to facilitate the collection and distribution of passengers [1]. Vehicles desiring to enter a DMV network from the manual highway will have to be processed through an entrance area where vehicle diagnostic checkout, destination selection, and route assignment will individually occur for each vehicle at entrance stations. After being processed, a vehicle will be accelerated from the entrance station it occupied and then merged into the high-speed (about 60 mph) guideway traffic.

In a typical DMV network, the number of vehicles desiring to pass through one entrance would commonly range up to 1000 per hour, and could occasionally approach double that amount [2]. The efficiency with which a large number of vehicles can be entered onto a link in a DMV network through an entrance configuration will directly affect network usage, and hence will have a significant effect on the feasibility of a DMV system. A previous paper [3] examined queuing characteristics of various DMV entrance configurations as a function of guideway occupancy just upstream from the entrance, but without consideration of vehicle flows elsewhere in the network. In [4], the basis for a system operating philosophy for a realistic urban DMV network was presented, and operation of an entrance to a DMV network was described under a strategy for preprogramming the routes of entering vehicles and their times of arrival at merge points in the network. The purpose of this paper is to examine the efficiency of an entrance to a DMV network when the entrance is operating in a manner consistent with the preprogramming strategy proposed in [4], and as a function of vehicle flows in the network. By means of a simulation model, the effects on entrance capacity and network utilization of parameters which are dependent on vehicle capabilities and DMV system design are studied.

NETWORK EFFECTS ON ENTRANCE OPERATION

The basic network management problem associated with operation of a DMV network is that smooth passage for all vehicles through the network must be achieved while high usage of network links is maintained and large queues of vehicles at entrances to the network are prevented. The need to accomplish the task of coordination of vehicle flows in a typical urban DMV network containing perhaps 25 interchanges (each interchange contains eight high-speed merge points) in the face of stochastic vehicle demands at entrances requires a strategy for preprogramming the route of each vehicle, and its time of arrival at each high-speed merge point on its route, before the vehicle is allowed to enter the DMV network. A preprogramming strategy potentially capable of accomplishing the necessary coordination of vehicle flows in a large DMV network containing interchanges was presented in [4]. In this section, a summary of the operation of an entrance to a DMV network using that preprogramming strategy will be given.

The fundamental philosophy for control of vehicles in a DMV network is the synchronously moving cell concept [5,6], under which it is assumed that a DMV network is "covered" by hypothetical contigous cells which move through the network in a predetermined, synchronous manner. A vehicle on the network tracks, or occupies, a particular cell until the system control assigns the vehicle to a different cell. The preprogramming strategy considers cells in groups called cycles, with a cycle defined as n contiguous cells. Because of the synchronous movement of the cells, the exact travel time for a cycle between points on the network is predetermined. In particular, the travel time is known for a cycle from a point near an entrance to every downstream "output point" in the network. (One output point exists on each outbound lane of every interchange. The output points are located a short distance downstream from the high-speed merge points in an interchange area and the locations of the output points depend on time synchronization requirements in the DMV network [4].) A computer keeps track of the number of vehicles already on the network which will be present in all future cycles at each output point, and the preprogramming strategy acts so as to schedule no more than n vehicles to pass any output point in the network in any one cycle. The cycle size n is chosen as the maximum number such that every interchange is capable of operating in a manner which assures smooth flow of vehicles into and out of the interchange, and hence smooth flow of vehicles everywhere on the network is possible.

Refer now to Fig. 1, which shows a portion of a DMV network containing two interchanges, an entrance A, and an exit E. Also shown is one of the four output points at each interchange, points C and D, respectively. Suppose that a vehicle at entrance point A wishes to travel through C and D to exit E. Before entering the network, the vehicle is assigned to a cell in a particular cycle that will soon pass merge point M. Travel times through the network are such that, no matter to which cell the vehicle is assigned in that cycle, the vehicle will be among the vehicles in the cycle which will begin to pass output point C at a specific future time, and among the vehicles in the cycle that will begin to pass output point D at another specific future time.

Thus, as far as scheduling of vehicles through the network is concerned, the particular cell that a vehicle is commanded to occupy in its assigned cycle does not matter. Hence, in the region on the guideway before an entrance merge point (the region from F to M in Fig. 1), vehicles on the guideway can be allowed to execute commands to shift between cells of their assigned cycle in order to make any particular cell of that cycle available at point M for occupancy by an entering vehicle. (Limitations in the ability of vehicles on the guideway to shift between cells in this manner will be discussed in the next section). The result is that the vehicle at entrance A which desires to travel through C and D to exit E can leave the entrance area for the guideway at the earliest time, determined by means of a computer check, such that 1) The cycle which the vehicle will be in when it reaches the guideway contains less than n vehicles; 2) A particular cycle that will pass output point C presently contains less than n vehicles; and 3) A particular cycle that will pass output point D presently contains less than n vehicles.

To generalize for any entrance to a DMV network, suppose a vehicle at an entrance wishes to travel through α output points at α interchanges ($\alpha = 0, 1, 2, \ldots$). Under the preprogramming strategy assumed, that vehicle can leave for the guideway when the cycle it will be in when it reaches the guideway contains less than n vehicles and cycles scheduled to pass the α output points at specific future times (which times depend on the fixed cycle travel times from the entrance to the respective output points) also presently contain less than n vehicles. Note that limitations on the capacity of exits are not being considered in this paper.

FORMULATION OF THE SIMULATION MODEL

The previous section discussed in general terms the procedure to be followed in deciding when a single vehicle waiting at an entrance can enter a network with assurance of smooth passage through the network. This section contains information on the entrance configurations that are considered in this paper and gives definitions and describes assumptions made in the formulation of the simulation model that is used in the study. The parameters are specified whose effects on entrance capacity and network utilization are to be examined, and the dependence of parameters on vehicle capabilities and system design is indicated.

The entrance configurations considered in this paper are of the type shown in Fig. 2. For convenience denote a configuration which has i stations in series and j stations in parallel as an "i x j" configuration. Define queue length in each of the j lanes as the number of waiting vehicles in that lane plus the number of stations ahead of the first vehicle in the lane. Vehicles are assumed to arrive from the external roadway system according to a Poisson distribution, and each arriving vehicle is assumed to choose to enter the lane with the shortest queue. For each set of parameters, 70 minutes of operation were simulated, and no data were collected for the first 10 minutes in order to allow initial transients to die out. The 70 minute simulation runs were repeated as necessary in order to reduce statistical fluctuations, and the results of the runs were averaged.

Define system time for a vehicle as the elapsed time from when the vehicle arrives in the entrance area to when that vehicle leaves for the guideway. The system time is made up of three parts: the wait time, the service time, and the post-service time. The wait time is defined as the length of time a vehicle waits in the entrance area before it can enter a station. The service time is defined as the time needed for the following operations: 1) positioning of the vehicle at a station, 2) vehicle diagnostic checkout as required, 3) switching to the automatic control mode, and 4) destination selection by the driver of the vehicle. The post-service time is defined as the time from when the four above operations have been completed to when the vehicle leaves a station as it begins accelerating towards the guideway.

The probability distribution of the service time obviously depends on system design. In this paper the service time probability density function is assumed to be of the form in Figure 3. The average value of the service time, $\bar{t}$, is one of the parameters in the study.

The post-service time for a vehicle depends on 1) the time needed for communication with the central computer, 2) the route of the vehicle, 3) vehicle flows in the network, and 4) interference beyond the stations between vehicles desiring to enter the guideway from one entrance area. The time needed for communication with the central computer will be neglected in this paper. The route of the vehicle and vehicle flows in the network will be treated in the next section. Interference beyond the stations in the entrance area between vehicles desiring to enter the guideway from one entrance area occurs because of the need to merge the j lanes in the entrance area (Figure 2) into the one-lane entrance ramp which merges with the main guideway. The interference between vehicles may require that successive vehicles heading for the guideway leave their respective stations at least τ seconds apart, with the value of τ depending on the entrance configuration, the construction of the ramp leading from the stations to the guideway, and the available trajectories for vehicles on the ramp. The effects of different values of τ are examined in this paper.

Throughout the simulations, the guideway cell velocity v_o and cell length L_C are assumed to be 88 ft/sec (60 mph) and 40 ft, respectively, resulting in a lane capacity of 7,920 vehicles per hour. In a network operating at 60 mph, the cycle size n is limited by the acceleration and deceleration capabilities of vehicles operating near 60 mph, the cell length, time synchronization requirements within an interchange area, and the allowable deviation, due to the need to satisfy safety requirements, from the usual vehicle velocity of 60 mph. It is expected that vehicle performance limitations will restrict the value of the parameter n to be $\leqslant 10$. The effects of values between 1 and 10 for this important parameter are examined in this study.

At this point, limitations in the ability of vehicles on the guideway to shift between cells in the region before an entrance merge point will be considered. Define the time necessary for passage of a cycle as T, and note that $T = nL_C/v_o$. Approximately, T = 3.18 sec when n=7 and T = 4.55 sec when n=10. Suppose that τ=2 sec, which means that successive vehicles must leave for the guideway at least 2 sec apart. Then, in the simulation, for n=7 no more than 2 vehicles can be entered into any one cycle at any one entrance, while for n=10 (the maximum cycle size considered) no more than 3 vehicles can be entered into any one cycle at any one entrance. This means that, if τ=2 sec, vehicles on the guideway in the region before an entrance merge point would be required to maneuver a maximum of 2 cells when n=7 and a maximum of 3 cells when n=10. This amount of maneuvering in the available space before an entrance merge point should be achievable. If necessary, vehicles can be made to occupy the most forward cells in a cycle before the cycle approaches the entrance merge point, so that if maneuvering is required it would be accomplished by slowing down the vehicles. This technique would effectively increase the amount of maneuvering that is possible, since much more maneuvering can be accomplished in a limited space by slowing down vehicles than by causing the vehicles to speed up. Thus, whenever τ is $\geqslant 2$ sec, the maneuvering required by the preprogramming strategy in the region on the guideway before an entrance merge point can very likely be accomplished.

APPROXIMATING THE NETWORK EFFECTS

The approach in this paper to approximating the effects on entrance efficiency of vehicle flows in the network is to approximate, in a manner to be described later in this section, the patterns of cycle occupancies "seen" by a vehicle waiting at an entrance. The pattern of cycle occupancies just upstream from the entrance is approximated as a function of average guideway occupancy at that point. The pattern of cycle occupancies at each output point on the route of the vehicle is approximated as a function of "average guideway occupancy at the output point as referred back to the entrance," a concept which is discussed in the next paragraph. In all cases, guideway occupancies are expressed in terms of percent of guideway lane capacity, e.g., 10% guideway occupancy just upstream from an entrance means that, on the average, 792 vehicles pass that point in an hour.

A vehicle waiting at an entrance competes for space in a cycle scheduled to pass an output point with those vehicles for which cells in the cycle have already been allocated at that output point. Such vehicles entered the network from an entrance farther upstream (in time) from the output point than the present entrance. Because of this effect, in this paper guideway occupancy at an output point is the "average occupancy at the output point as referred back to the entrance" whose operation is being studied. For example, Fig. 4 shows travel demands on a typical guideway network (from [2]), an output point A, and entrances B and C. Fig. 5 shows the same network and indicates that, according to the travel demands in Fig. 4, average guideway occupancy at A is 81% during a typical rush hour. Also shown in Fig. 5 are lines of equal travel time to A from guideway locations upstream from A, and the average guideway occupancies at A as referred back to upstream locations for the trip pattern which led to Fig. 4. That is, Fig. 5 indicates that on the average over the rush hour 5% of lane capacity at A is allocated to vehicles which must travel more than 15 minutes on the network before reaching A. Therefore, the average guideway occupancy at A as referred back to entrance B, which entrance is slightly less than 15 minutes from A, is 5%. On the other hand, the average guideway occupancy at A as referred back to entrance C is 60%, since on the average 60% of lane capacity at A is allocated to vehicles which are already on the network when vehicles entering at C request entrance to the network. Any vehicle entering at C and wishing to be routed through A on the average "sees" a certain guideway percent occupancy just upstream from C, a guideway occupancy of 60% at output point A, and certain guideway percent occupancies at other output points on its route.

For simplicity in the simulation, it is assumed that all vehicles entering at an entrance wish to pass through the same output points. In an actual network, of course, this is not the case. However, in an actual network guideway occupancy at an output point as referred back to an entrance generally is below 50% for all but a few output points, and as will be seen in the simulation results, guideway occupancies below about 50% have a negligible effect on entrance capacity for reasonable cycle sizes. Hence, the capacity of any one particular entrance in a network would be significantly influenced by the guideway occupancies, as referred back to that entrance, at only a few output points in the network. Therefore, although an exact comparison of the simulation results with capacity of an entrance in an actual network cannot generally be made, it is expected that use of results obtained by this type of simulation would permit a good estimate to be made of what actual entrance capacity would be in a network.

The method used in the simulation for generating cycle occupancies "seen" by vehicles desiring to enter at a particular entrance (call it entrance E) will now be described. The idea is to approximate the random patterns of cycle occupancies that occur in a network due to the fact that vehicles randomly arrive at many entrances in the network and reserve space in cycles which at specific future times will pass entrance E and the output points downstream from E. The pattern of cycle occupancy just upstream from

E and the patterns of cycle occupancy for each of the output points being considered are generated independently of each other.

Given an average guideway occupancy of y% when guideway lane capacity is x vehicles per hour, the method for generating cycle occupancies assumes that N entrance stations, where N = 0.001 x y, each independently have a probability p of attempting to enter exactly one vehicle into a cycle and a probability q = 1 - p of entering no vehicles into that cycle. The probability p is chosen such that, on the average, each entrance station enters 10 vehicles in an hour, i.e., p = 10T/3600, where T is the cycle time in seconds, as defined previously. The N entrance stations collectively have the probability P(i), where

$$P(i) = \binom{N}{i} p^{i} q^{(N-i)}, \qquad (1)$$

of entering i vehicles in any given cycle. The occupancy of each successive cycle within the simulation time is generated sequentially, and any vehicles in excess of n vehicles in a given cycle are placed into the next cycle which is not already occupied by n vehicles.

This method for generating the pattern of cycle occupancies for the simulation leads to the bunching of fully occupied cycles, an effect which can be expected to naturally occur in a network where stochastic vehicle demands are present at entrances. This bunching effect is worse in the simulation than it actually would be in a network, since the simulation does not represent the exiting of vehicles which occurs in a network and which tends to reduce the number of full or nearly full cycles. Because of this fact, and since in the simulation the patterns of cycle occupancy for output points are obtained independently of each other and independently of the pattern of cycle occupancy just upstream from the entrance being considered, the simulation results give a somewhat pessimistic estimate of entrance capacity for a given choice of parameter values.

SIMULATION RESULTS

The effects of the parameters on entrance capacity are examined in this paper by comparing, for sets of parameter values, the maximum number of arrivals per hour yielding an average system time per arriving vehicle of 30 seconds. The number of arrivals per hour obtained using this criterion will be denoted as the effective capacity of a given entrance configuration with the assumed parameter values. A typical curve of average system time per arriving vehicle vs the number of vehicles arriving per hour, as shown in Figure 6, indicates that the 30 sec criterion yields an estimate of entrance capacity which is less than maximum capacity with a given set of parameters. The Figure also shows that, with an average system time per arriving vehicle of 30 sec, the system time for 90% of the arriving vehicles is less than about 60 sec, i.e., approximately 10% of the arriving vehicles are in the entrance area for over one minute when average system time per arriving vehicle is 30 sec.

The effects of cycle size will be discussed with the aid of Table 1. A 1x5 entrance configuration is assumed, and the values of the parameters which are held constant are indicated at the bottom of the Table. For convenience, the form $X_o/X_1/X_2$ is used to denote the assumed percent occupancy of the guideway just upstream from the entrance (X_o), and the assumed percent occupancy at each of two output points as referred back to the entrance (X_1 and X_2, respectively). Thus, for example, with 60% occupancy just upstream from the entrance and 50% occupancy at output point 1 as referred back to the entrance (the 60/50/0 case), the effective capacity of the entrance configuration is 1010 vehicles per hour when cycle size equals 10.

The data in Table 1 allow several important observations to be made about the effects on entrance capacity and network utilization of the number of cells per cycle. First of all, part A of Table 1 indicates that cycle size has little influence on the effective capacity of an entrance configuration when none of the entering vehicles must be routed through output points, i.e., there are no high-speed merges downstream from the entrance. However, note in parts B and C of the Table the effect that the number of cells per cycle has on the effective capacity of an entrance when entering vehicles must be routed through output points, as is the usual case in a network. For example, for the 70/70/70 case, the effective capacity of the entrance configuration is zero veh/hr when n=1, indicating not only that entrance efficiency is severely degraded, but also that network usage is poor, since essentially no vehicles could enter the network under those specified network conditions. On the other hand, 500 veh/hr can be entered under the same network conditions when n=10. Effective entrance capacity and network usage are thus both significantly improved by increasing the number of cells per cycle from 1 to 10.

TABLE 1
EFFECTIVE CAPACITY AS A FUNCTION OF CYCLE SIZE

Percent Occupancy				Number of Cells per Cycle (n)				
X_o	X_1	X_2		10	7	5	3	1
0	0	0		1100	1100	1100	1100	1100
50	0	0		1080	1060	1040	1030	980
60	0	0	PART A	1040	1020	--	--	940
70	0	0		925	900	--	--	810
80	0	0		660	610	--	--	560
90	0	0		125	80	--	--	--
50	50	0		1050	1030	1000	940	850
60	50	0		1010	990	--	--	750
60	60	0		1000	910	--	--	--
70	70	0	PART B	725	630	570	490	320
80	60	0		590	480	--	--	--
80	80	0		300	140	--	--	--
50	50	50		1030	990	940	835	570
70	70	70		500	360	270	150	0
80	70	60	PART C	390	230	--	--	--
80	70	70		250	--	--	--	--
80	80	70		60	--	--	--	--

Configuration: 1 x 5 $\bar{t} = 15$ sec
Lane Capacity = 7920 veh/hr $\tau = 0$ sec

Another important observation can be made by plotting some of the data in Table 1 in the form shown in Figure 7. By comparing the curves in the Figure it can be observed that guideway occupancies less than 50% have a negligible influence on effective entrance capacity for $n \geqslant 7$, while for $n < 5$ guideway occupancies of 50% do have a significant influence. Hence, if n=7, the results indicate that there would usually be little difficulty in scheduling a vehicle through those output points on its route for which guideway occupancy is less than 50%. If instead n=3, there would generally be difficulty in scheduling a vehicle through output points with occupancies less than 50%. These observations indicate the critical importance to entrance efficiency and network utilization of achieving at least 5 and preferably 7 cells per cycle.

The effective capacity of a 1x5 entrance configuration as a function of average service time $\overline{t}$ is shown in Table 2. Since typical entrance volumes would commonly range up to 1000 vehicles per hour, the data in Table 2 indicate a need to minimize the time needed to complete the operations which make up the service time. Otherwise, in the congested portions of the network large entrance configurations would be required to handle the number of vehicles desiring to enter the network.

TABLE 2

EFFECTIVE CAPACITY AS A FUNCTION OF AVERAGE SERVICE TIME

Percent Occupancy			Average Service Time $(\overline{t})$		
X_o	X_1	X_2	15 Sec	20 Sec	25 Sec
50	0	0	1080	750	525
60	0	0	1040	710	470
70	0	0	925	640	440
80	0	0	660	475	230
50	50	0	1050	--	510
60	50	0	1010	--	465
60	60	0	1000	--	460
70	70	0	725	--	325
80	60	0	590	--	210
80	80	0	300	--	125

Configuration: 1x5 n=10 cells per cycle
Lane Capacity = 7920 veh/hr τ=0 sec

In Table 3, the effect of interference between vehicles in the entrance area is examined by means of the parameter τ, which specifies the minimum allowable difference between the departure times of successive vehicles that leave their respective stations and head for the guideway. Note from the Table that increasing τ from 0 to 2 sec causes a greater reduction in effective entrance capacity when cycle size is one than when cycle size is 7. Once again the importance of cycle size is evident. It is also evident from the Table that, for a cycle size of 7, an increase in τ from 0 to 2 sec has a minor influence on entrance capacity. Of course,

TABLE 3

EFFECTIVE CAPACITY AS A FUNCTION OF τ

Percent Occupancy			7 Cells Per Cycle τ(Sec)			1 Cell Per Cycle τ(Sec)	
X_o	X_1	X_2	0	2	3	0	2
70	0	0	900	815	710	810	650
80	0	0	610	530	455	560	370
70	70	0	630	570	495	320	220
50	50	50	990	900	805	570	420

Configuration: 1x5
Lane Capacity = 7920 veh/hr $\overline{t}$=15 sec

the relative effect of τ on entrance capacity increases as τ increases to 3 sec, which for the configuration considered is the average processing time per vehicle divided by the number of stations.

Table 4 contains comparisons between the effective capacities of several entrance configurations. Note from the Table that, for guideway occupancies greater than 60%, a 2x3 configuration containing 6 stations is less efficient that a 1x5 configuration containing 5 stations. By comparing the data for the 1x5 and 2x5 configurations (n=7 cells per cycle), it can be seen that the effective capacities of the two configurations are quite low and only slightly different for guideway occupancies like 90/0/0 and 80/80/0. That is, doubling the number of entrance stations results in only a small increase in effective entrance capacity for those guideway occupancies. However, the data in the Table also indicate that even for a high-occupancy case like 80/60/0, enough vehicles can be entered through an entrance configuration to make the guideway occupancy approach 90% at some point in the network. Thus, with a cycle size of 7, a guideway occupancy of 90% is approximately the maximum occupancy achievable anywhere in a network by use of the preprogramming strategy presented in [4] and considered in this paper.

TABLE 4
EFFECTIVE CAPACITIES OF DIFFERENT ENTRANCE CONFIGURATIONS

Percent Occupancy			10 Cells Per Cycle		7 Cells Per Cycle	
X_o	X_1	X_2	1x5	2x3	1x5	2x5
50	0	0	1080	1125	--	--
60	0	0	1040	1010	--	--
70	0	0	925	880	900	1400
80	0	0	660	570	610	880
90	0	0	125	100	80	110
70	70	0	--	--	630	830
80	60	0	--	--	480	600
80	80	0	--	--	140	150
70	70	70	--	--	360	390

Lane Capacity = 7920 veh/hr $\quad \bar{t}=15$ sec
$\tau=0$ sec

SUMMARY AND RECOMMENDATIONS

This paper examined DMV entrance capacity in a manner consistent with a chosen system operating philosophy and as a function of vehicle flows in a DMV network. The effects on entrance capacity and network utilization of various parameters whose values are dependent on vehicle capabilities and system design were studied.

The results of the study indicate the importance to entrance capacity and network utilization of designing a DMV system so as to achieve a minimum cycle size of 5. With a cycle size of 7, network utilization would be improved, and the approximate maximum occupancy achievable on the busiest link in a network would be 90%.

Consideration must be given to reducing the number of operations each vehicle must undergo while being processed through an entrance area, since the amount of time needed for these operations (the service time) has a significant influence on entrance capacity. With a cycle size of 10, for instance, an average service time per vehicle of 25 seconds would require large entrance configurations to handle typical entering volumes in congested portions of a network.

The design of the acceleration region between the stations in the entrance area and the one-lane entrance ramp leading to the guideway is important for high-volume entrances. System design should assure that successive vehicles can leave their respective stations and head for the guideway at time intervals of about 2 sec or less. Obviously, the length of the time interval puts an absolute limit on entrance capacity. Therefore, early attention should be given to the problem of how the low speed merges that are necessary in the acceleration region can be accomplished while the time interval is kept small.

As the design of a DMV system evolves, better information about parameter values will be obtained than is available at present. Further studies of entrance capacity will be required. In particular, the number of parameter values that had to be considered in the present paper allowed only a quick look to be taken at the differences between the capacities of various entrance configurations. The lane capacity for a DMV system may be different than that assumed here. Finally, information about exit design must be obtained so that if any exit volume restrictions exist they can be incorporated into the study of entrance capacities.

ACKNOWLEDGEMENT

The author would like to acknowledge the work of Mr. Stephen J. Kiselewich who participated in discussions and wrote the computer programs for this study.

REFERENCES

1. System Analysis of Urban Transportation, General Research Corporation, conducted under contract H-777, U. S. Department of Housing and Urban Development, January, 1968.

2. R. G. Stefanek and D. F. Wilkie, "The Impact of a Dual-Mode Vehicle System on Transportation in the Detroit Area," Report 70-22, Transportation Research and Planning Office, Ford Motor Company, Dearborn, Michigan, October, 1970.

3. S. J. Kiselewich, "Queuing Characteristics of Several Automated Highway Entrance Ramp Configurations," Report 70-17, Transportation Research and Planning Office, Ford Motor Company, Dearborn, Mich., September, 1970.

4. R. G. Stefanek, "Network Implications on Control System Design for a Dual-Mode Transportation System," Report 71-4, Transportation Research and Planning Office, Ford Motor Company, Dearborn, Mich., April, 1971.

5. Study of Synchronous Longitudinal Guidance as Applied to Intercity Automated Highway Networks, TRW Systems, Inc., prepared under contract to U. S. Department of Transportation, contract no. C-353-66, Sept. 1969.

6. D. F. Wilkie, "A Moving Cell Control Scheme for Automated Transportation Systems," Transportation Science, vol. 4, no. 4, November, 1970.

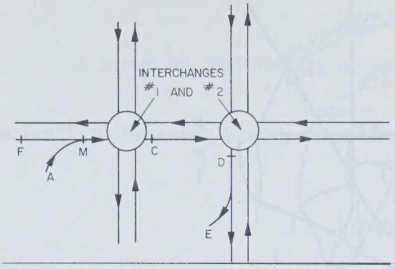

INTERCHANGES
#1 AND #2
F
M
C
A
D
E

Fig. 1 Portion of a DMV Network

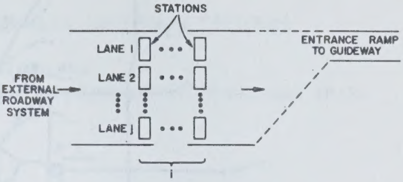

STATIONS
LANE 1
LANE 2
LANE j
FROM EXTERNAL ROADWAY SYSTEM
ENTRANCE RAMP TO GUIDEWAY
i

Fig. 2 DMV Entrance Area

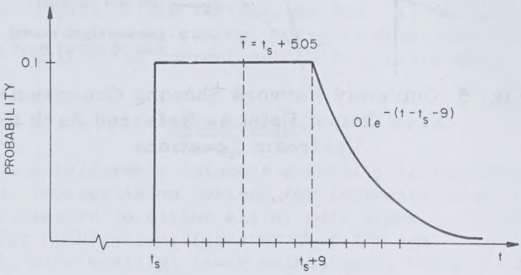


Fig. 3 Service Time Probability Density Function

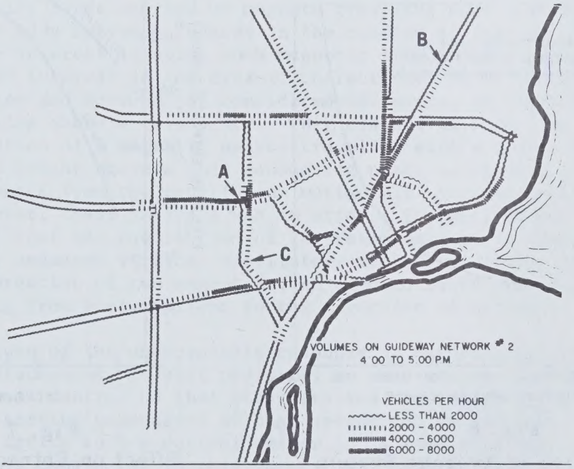

A
B
C
VOLUMES ON GUIDEWAY NETWORK # 2
4:00 TO 5:00 PM
VEHICLES PER HOUR
LESS THAN 2000
2000 - 4000
4000 - 6000
6000 - 8000

Fig. 4 Volumes on a Typical Guideway Network

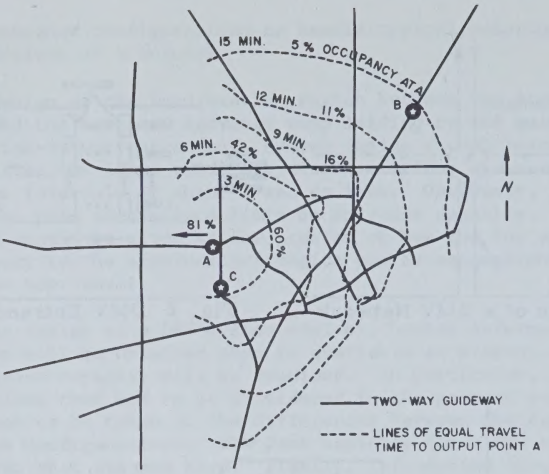

15 MIN.
5 % OCCUPANCY AT A
12 MIN.
11%
B
9 MIN.
6 MIN.
42%
16%
3 MIN
N
81%
60%
A
C
TWO-WAY GUIDEWAY
LINES OF EQUAL TRAVEL TIME TO OUTPUT POINT A

Fig. 5 Guideway Network Showing Occupancies at an Output Point as Referred Back to Upstream Locations

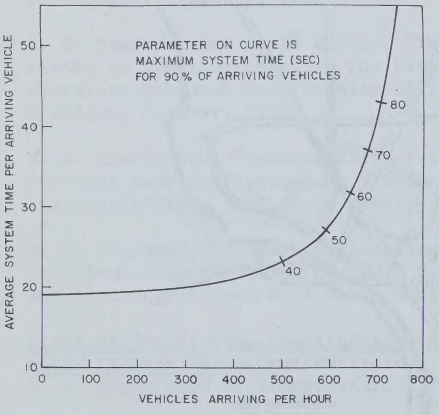


Fig. 6

Typical Curve of Average System Time per Arriving Vehicle vs. Vehicles Arriving per Hour

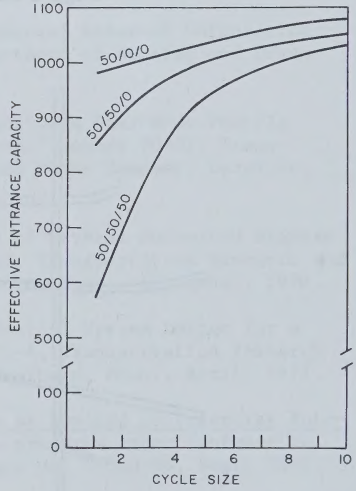


Fig. 7

Effect on Entrance Capacity of 50% Occupancy at Additional Output Points

A STUDY OF THE CONTROL OF MAGNETICALLY LEVITATED VEHICLES

Dennis F. Wilkie
Scientific Research Staff, Ford Motor Company, Dearborn, Michigan 48121

ABSTRACT

One of the methods of suspension proposed for high speed ground vehicles is magnetic levitation. However, the motion of levitated magnets which are perturbed (vertically) from equilibrium is oscillatory with very little damping. In this paper, a feedback control system is used to obtain sufficient damping of the vertical motion. It is found that very good damping can be obtained by control currents which are a small percentage of the equilibrium current needed for levitation.

I. INTRODUCTION

There is considerable interest currently in the development of high speed ground transportation systems for intercity travel which offer an attractive alternative to either air or auto travel. In the speed range desirable for such systems (greater than 250 mph), problems such as force transmittal, ride quality, track maintenance, vehicle wear, etc. associated with conventional steel wheel on steel rail vehicle suspension systems are presently so formidable that other means of vehicle support must be investigated. The alternate schemes that have been proposed for vehicle support are air cushions [1] and systems of magnetic levitation [1-7]. One of the most promising approaches to magnetic levitation utilizes the lift force exerted on magnets traveling over a conducting roadbed by the eddy currents induced in the roadbed to support the vehicle. Because of the interest in using such magnetic levitation schemes, there has been recent interest in the dynamic characteristics of levitated magnets. Bailey and Norwood [8] considered the motion of various magnetic sources traveling above a thin plate. They claimed to show that the equilibrium motion of a magnetic source traveling with a fixed velocity and at a fixed height above a thin conducting plate is stable, i.e., that small disturbances from the equilibrium motion will asymptotically decay to zero. However, their analysis was in error. In fact, it has recently been shown [7] that the equilibrium of levitated magnets is unstable with an essentially undamped vertical oscillatory motion resulting from vertical perturbation of the magnet and an exponentially growing velocity error resulting from perturbations in the direction of motion.

Because of the unacceptable response of magnetic levitation systems to disturbances, it will obviously be necessary to implement a means of feedback control so that stability and ride characteristics suitable for carrying passengers at high speeds can be achieved. The problem of designing such a control system for a realistic magnet configuration is considered in this paper. In Section II, the results of [7] regarding approximate force laws for magnets in non-steady motion over a conducting plate are briefly reviewed. In Section III, the motion of a current carrying coil over a conducting plate is analyzed, and it is seen that the horizontal and vertical motion are essentially decoupled. Based on the results of this analysis, a feedback current controller is designed by a linear (small signal) analysis of the dynamic equations of the coil in Section IV, and the motion of the controlled magnet system is analyzed. It is seen that relatively small currents are required to obtain acceptable damping of the vertical motion. In order to validate these theoretical predictions, a small scale experiment with feedback control on a levitated coil was built. The results obtained experimentally agree very well with the theoretical analysis, and they are presented in Section V. In Section VI, alternate means of implementing the feedback control are discussed since there might be practical difficulties in implementing a control directly on the current in the main levitation coils (e.g., if the levitation coils use superconductors in the persistent mode). It is seen that the effect of the control coils falls off rapidly as the coils are moved away from a location co-axial with the main levitation coils. Section VII contains the conclusions.

II. THE FORCE LAWS FOR MAGNETIC LEVITATION

The published results regarding the forces involved in magnetic levitation are reviewed in this section for the sake of completeness.

Reitz [5] has obtained expressions for the lift and drag forces of various magnet configurations in steady motion over an infinite thin conducting plate. For example, in the case of an infinitely long current carrying wire, the lift force per unit length on the wire was found to be

$$F_L = \left(\frac{\mu_o I^2}{4\pi h} \right) \left[\frac{v^2}{v^2 + w^2} \right] \tag{2.1}$$

where

F_L = lift force in newtons
I = current in the wire
μ_o = permeability of free space ($4\pi \times 10^{-7}$ Wb./amp. m)
h = constant height above the conducting plate (meter)
v = the horizontal constant velocity of the wire (meter/sec.)

and w is a constant with the units of velocity given by

$$w = \frac{2}{\mu_o \sigma \delta} \tag{2.2}$$

where σ is the conductivity of the conducting plate and δ its thickness. The drag force experienced by the wire in the horizontal direction was found to be given by the expression

$$F_D = \left(\frac{w}{v} \right) F_L \ . \tag{2.3}$$

This drag is due to the eddy current losses in the thin conducting plate. It was found in [5] that the relation (2.3) holds for all magnet configurations of interest. Reitz [5] also analyzed the case of a current carrying coil over a conducting plate, but a simple analytical expression similar to (2.1) cannot be derived for that case. Numerical values for the lift force on a coil are given in [5]. It is found that the following empirical relation fits that data very well,

$$F_L = F_I \left[1 - \frac{1}{(1 + v^2/w^2)^n} \right] \tag{2.4}$$

where F_I is the repulsive force between the current carrying coil and its image located at an equal distance below the conducting plane. The parameter n depends on the coil geometry and h. For a 1 meter x 2 meter coil (2 m. in the direction of motion) at a nominal height of h = .1m, $n \cong \frac{1}{4}$. An analytical expression for the ideal lift force on a current carrying coil is given in the next section.

It would seem that a good approximation to the force laws for a magnet deviating only slightly from steady motion over a thin conducting plate would be obtained by replacing the velocity, v, and height, h, in (2.3) and (2.4) with the instantaneous velocity and height of the magnet. Referring to the coordinate system of Figure 1, for the infinite wire, the lift force on the wire deviating a small amount from steady motion would be

$$F_L(t) = \left(\frac{\mu_o I^2}{4\pi z(t)} \right) \left[\frac{\dot{x}(t)^2}{(\dot{x}(t))^2 + w^2} \right] \tag{2.5}$$

where $\dot{x}(t) \equiv dx(t)/dt$. Davis and Wilkie [7] have investigated the validity of this approximation, referred to as the instantaneous approximation. By comparing the motion obtained by a more exact analysis of various magnet systems with that given by the forces resulting from the instantaneous approximation, they have shown that this approximation is indeed very good for all cases of interest. This approximation greatly facilitates the dynamical analysis of practical magnetic levitation systems since the exact equations of motion cannot be solved analytically and are not even amenable to numerical solution. The drag force on a magnet system would then be obtained by replacing F_L and v in (2.3) with $F_L(t)$ in (2.5) and $\dot{x}(t)$ respectively.

The other results in [7] regarding the motion of levitated magnets which are of significance here are the following:

(1) The horizontal and vertical motion resulting from perturbations to steady motion are essentially uncoupled.

(2) The vertical motion resulting from perturbations is an oscillatory motion with a very small amount of damping. (The damping can be either positive or negative depending on the magnet configuration.)

(3) With constant propulsive force, any perturbation to horizontal motion will result in velocity errors which grow exponentially.

Because of these results, the need for some means of feedback control to obtain acceptable responses to perturbations for practical magnet configurations is very apparent.

III. DYNAMICS OF MOTION OF A LEVITATED COIL

In this section, the dynamics of motion of a current carrying coil which is perturbed from steady motion at a velocity v and height h over an infinite, thin conducting plate are analyzed. The force laws for this motion are obtained from the results in [5] using the instantaneous approximation discussed in [7]. The coordinate system for the motion of the coil is shown in Figure 2. The coil is assumed to be traveling in the plus x direction (indicated by arrow) in the plane of the paper over the conducting x-y plane. The nominal velocity in the x direction is constant (v) as is the nominal height above the xy plane (h).

An analytical expression for the magnetic lift and drag forces on a coil experiencing steady motion are not given explicitly by Reitz [5]. However, numerical values of the forces are given and an analytical expression for the ideal lift force between the coil and its image located directly below the coil at a distance h below the x-y plane can be derived using the mutual inductance expression in [5]. Thus, F_I and n in (2.4) can be determined for the coil.

The mutual inductance between a coil and its image directly below it is given in [5] as

$$M(\hat{z}) = \frac{\mu_o N^2}{2\pi} \Big\{ 2a \sinh^{-1}\left(\frac{a}{\hat{z}}\right) - 2a \sinh^{-1}\left(\frac{a}{\sqrt{b^2 + \hat{z}^2}}\right)$$

$$+ 2b \sinh^{-1}\left(\frac{b}{\hat{z}}\right) - 2b \sinh^{-1}\left(\frac{b}{\sqrt{a^2 + \hat{z}^2}}\right) \tag{3.1}$$

$$+ 4\sqrt{a^2 + b^2 + \hat{z}^2} - 4\sqrt{a^2 + \hat{z}^2} - 4\sqrt{b^2 + \hat{z}^2} + 4\hat{z} \Big\}$$

where

N = number of turns in the coil

$\hat{z}$ = 2h = distance between the coils

a = width of coil

b = length of coil (in direction of motion).

The lift force exerted on the coil due to this single image is given by

$$F_i = - I^2 \frac{dM(\hat{z})}{d\hat{z}} , \tag{3.2}$$

and differentiation of (3.1) yields

$$\frac{dM}{d\hat{z}} = \frac{\mu_o N^2}{2\pi} \left\{ - \frac{2a^2}{\hat{z}\sqrt{a^2 + \hat{z}^2}} + \frac{2a^2\,\hat{z}}{(b^2 + \hat{z}^2)\sqrt{a^2 + b^2 + \hat{z}^2}} \right.$$

$$- \frac{2b^2}{\hat{z}\sqrt{b^2 + \hat{z}^2}} + \frac{2b^2\,\hat{z}}{(a^2 + \hat{z}^2)\sqrt{a^2 + b^2 + \hat{z}^2}} \tag{3.3}$$

$$\left. + \frac{4\,\hat{z}}{\sqrt{a^2 + b^2 + \hat{z}^2}} - \frac{4\,\hat{z}}{\sqrt{a^2 + \hat{z}^2}} - \frac{4\,\hat{z}}{\sqrt{b^2 + \hat{z}^2}} + 4 \right\}$$

where a, b, $\hat{z}$, and N are as defined in (3.1). In order to obtain the equations of motion, let us define the following variables. Let

$$x_1(t) = z(t) - h = \text{deviation of height from nominal} \tag{3.4}$$

and let

$$\dot{x}_1(t) = \frac{dx_1}{dt}(t) = x_2(t) = \dot{z}(t) = \text{vertical velocity.} \tag{3.5}$$

Then

$$\dot{x}_2(t) = \ddot{z}(t) = \frac{1}{m}(F_z - mg) = \text{vertical acceleration.} \tag{3.6}$$

Using (3.2) and (2.4), along with (3.5) and (3.6), we obtain the following coupled first order differential equations which describe the vertical motion of the coil above the plane:

$$\dot{x}_1(t) = x_2(t) \tag{3.7a}$$

$$\dot{x}_2(t) = - \frac{I^2}{m}\left[1 - \frac{1}{\left(1 + \frac{v^2}{w^2}\right)^n}\right]\frac{dM}{d\hat{z}} - g\,. \tag{3.7b}$$

In this form, the equations of motion are known as state equations and $x_1(t)$ and $x_2(t)$ are state variables. In (7), it was shown that the coupling between the horizontal and vertical motion of a levitated wire was negligible. Indeed, the same decoupling can be shown (either analytically or numerically) to exist here. Thus, the dynamics of the vertical motion can be studied using (3.7) with the velocity v considered to be a constant. The current I in the present analysis is considered as constant at the equilibrium value I_e needed to support the mass m at a height h, i.e.,

$$I_e^{\,2} = \frac{-\,mg}{\left.\frac{dM}{d\hat{z}}\right|_{\hat{z} = 2h}} \tag{3.8}$$

Consider the case of motion which is deviating only slightly from the desired steady motion so that $x_1(t)$ and $x_2(t)$ are very small. Then the motion of the magnet should be quite well described by the linear approximations of (3.7) which are

$$\dot{x}_1(t) = x_2(t) \; , \tag{3.9a}$$

$$\dot{x}_2(t) = \frac{-2\, I_e^2}{m} \left(1 - \frac{1}{\left(1 + \frac{v^2}{w^2}\right)^n}\right) \cdot \left(\frac{d^2M}{d\hat{z}^2}\right)_{\hat{z} = 2h} \cdot x_1(t) . \tag{3.9b}$$

These equations may be simply written in matrix form as

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\alpha & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{3.10}$$

where

$$\alpha = \frac{2I_e^2}{m}\left(1 - \frac{1}{(1 + v^2/w^2)^n}\right) \frac{d^2M}{d\hat{z}^2}\bigg|_{\hat{z} = 2h}$$

It is easily shown that the form of the solution of (3.10) for non-zero initial conditions is

$$x_i(t) = a_i \cos\left(\sqrt{\alpha}\;\; t + \varphi_i\right), \; i = 1, 2 \tag{3.11}$$

whenever $\alpha > 0$. Utilizing the equilibrium condition in (3.8), an alternate expression for α can be shown to be

$$\alpha = -\left\{2g\left[1 - \frac{1}{(1 + v^2/w^2)^n}\right] \frac{d^2M}{d\hat{z}^2}\right\} \Big/ \frac{dM}{d\hat{z}} \; . \tag{3.12}$$

From the expressions for $\frac{dM}{d\hat{z}}$ and $\frac{d^2M}{d\hat{z}^2}$ it can be seen that in all cases here $\frac{dM}{d\hat{z}} < 0$ and $\frac{d^2M}{d\hat{z}^2} > 0$. Thus, the vertical response of the levitated coil to vertical disturbances is, to the first approximation, always an oscillatory, undamped motion with the frequency of oscillation given by $\sqrt{\alpha}$. It would be expected that the exact response to (3.7) would have a small degree of damping but certainly not enough to produce suitable responses to disturbances. As an example calculation of the oscillation frequency, consider the case of a 2m. x 2m. coil traveling at 134 m./sec. (300 mph) over a thin aluminum plate ($w = 4.5$ m/s). It can be shown that in this case $\frac{dM}{d\hat{z}} = -36$, $\frac{d^2M}{d\hat{z}^2} = 199$ and thus that the frequency of oscillation is $= \sqrt{\alpha}/2\pi = 1.67$ hertz. This is very close to the frequency found in [7] for the vertical motion of an infinite current carrying wire. Certainly such a response is unacceptable, and in the next section a simple feedback control scheme to obtain a suitably damped response is discussed.

IV. FEEDBACK CONTROL OF THE VERTICAL MOTION

It was shown in the previous sections that the vertical motion of a levitated coil resulting from disturbances is unacceptable. In this section, a feedback control policy for varying the current in the coil to damp out disturbances is analyzed, and it is seen that control currents which are very small compared to the equilibrium current can be used to obtain acceptable damping.

Assume the same state variable definitions as given by (3.4) - (3.6) previously. Assume further that for small deviations of the motion from equilibrium the current in the coil will be varied a small amount from the equilibrium value, i.e.,

$$I = I_e + i(t) \tag{4.1}$$

where $|i(t)| << I_e$. Then, if the equations of motion (3.7) are linearized about the nominal equilibrium point $x_1(t) = x_2(t) = i(t) = 0$, the resulting equations are

$$\dot{x}_1(t) = x_2(t) \tag{4.2a}$$

$$\dot{x}_2(t) = -\left(\frac{2I_e^2}{m} \cdot \eta \cdot \left.\frac{d^2M}{d\hat{z}^2}\right|_{2h}\right) \cdot x_1(t) - \frac{2I_e}{m} \cdot \eta \cdot \left.\frac{dM}{d\hat{z}}\right|_{2h} \cdot i(t), \tag{4.2b}$$

where $\eta \equiv \left(1 - \frac{1}{(1 + v^2/w^2)^n}\right)$.

Let us assume that the feedback control is such that the current is proportional to the vertical velocity,

$$i(t) = \hbar x_2(t) \ . \tag{4.3}$$

Then (4.2) can be written in the matrix form

$$\dot{\underset{\approx}{x}}(t) = \underset{\sim}{A}(t) = \begin{bmatrix} 0 & 1 \\ -\alpha & -\beta \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{4.4}$$

where

$$\alpha \equiv -\ 2g\ \eta \left.\frac{d^2M}{d\hat{z}^2}\right|_{2h} \Bigg/ \left. dM/d\hat{z}\right|_{2h}$$

and

$$\beta \equiv 2\,\frac{\hbar \eta I_e}{m} \left.\frac{dM}{d\hat{z}}\right|_{2h}$$

Using any one of a number of standard techniques, it can be shown that the solution of (4.4) with an initial position error x_1^o is given by

$$x_1(t) = \sqrt{\frac{\alpha}{\alpha-\beta^2/4}} \cdot x_1^o e^{-\beta t/2} \sin(\omega t + \varphi) \tag{4.5a}$$

$$x_2(t) = \frac{\alpha \cdot x_1^o}{\sqrt{\alpha-\beta^2/4}} e^{-\beta t/2} \sin(\omega t) \tag{4.5b}$$

$$\omega = \sqrt{\alpha-\beta^2/4} \quad , \quad \varphi = \tan^{-1}\left(\frac{2\sqrt{\alpha-\beta^2/4}}{\beta}\right)$$

For $\alpha > \beta^2/4$ (which holds in all cases of interest). Thus, the damping of the oscillatory response is directly controlled by varying β which is proportional to the hypothesized feedback gain $\hbar$. Thus, using the design rule that all transients in a damped system are zero after five time constants ($e^{-5} = .0067$), the feedback needed to damp out disturbances in T sec. is determined by

$$\beta = \frac{10}{T} \; . \tag{4.6}$$

An important question to answer is how much control current is needed for a given amount of damping. Obviously i(t) will be oscillatory and decaying exponentially with the same time constant as the terms in (4.5). Thus, let us determine the maximum value i(t) will take on for a given disturbance. Obviously,

$$|i(t)|_{max} = |\hbar x_2(t)|_{max} = |\hbar| \; |x_2(t)|_{max} \; . \tag{4.7}$$

Further, $|x_2(t)|$ takes on its maximum value at the time t_{max}, where

$$t_{max} = \frac{1}{\sqrt{\alpha-\beta^2/4}} \tan^{-1}\left[-\frac{2\sqrt{\alpha-\beta^2/4}}{\beta}\right] , \tag{4.8}$$

and thus

$$|x_2(t)|_{max} = \frac{x_1^o \alpha}{\sqrt{\alpha-\beta^2/4}} \exp\left[-\frac{\beta t_{max}}{2}\right] . \tag{4.9}$$

From the earlier definition of β, it follows that

$$|\hbar| = \beta m / (2 I_e \left|\frac{dM}{d\hat{z}}\right|_{2h}) = \left(\frac{\beta}{2}\right) \cdot \sqrt{\frac{m}{g \left|\frac{dM}{d\hat{z}}\right|_{2h}}} \; . \tag{4.10}$$

Thus $|i(t)|_{max}$ can be determined using (4.7), (4.9), and (4.10) for a given time for damping (i.e., a given β). Using (4.9), (4.10), and the expression for the equilibrium current given in (3.8), it follows that the ratio of the maximum control current to the equilibrium current is given by

$$\frac{|i(t)|_{max}}{I_e} = \frac{\beta}{2g} \cdot \frac{x_1^{o}\alpha}{\sqrt{\alpha-\beta^2/4}} \exp\left[-\frac{\beta t_{max}}{2}\right], \tag{4.11}$$

which is independent of the mass of the coil.

For an example calculation, consider the case of a 2m. x 2m. coil of 4,000 turns weighing 30 kg (commercially available superconducting wire). The equilibrium current needed to levitate such a coil at .1m above a plate at 134 m/sec. would 1.6 amperes. If a disturbance of 1 inch (25% of equilibrium height) is to be forced to zero in 5 seconds, the resulting value of $|i(t)|_{max}$ is found to be 36.5 ma or 2.28% of the equilibrium current. If this coil is to support the weight of a 50,000 kg vehicle, the values for I_e and $|i(t)|_{max}$ increase to 65.5 amperes and 1.49 amperes, respectively, but the percentage remains the same. Similar calculation of the peak current needed to damp out the oscillations in 1 second show that about 7.3% of the equilibrium current value would be needed for control. Thus, very good transient responses should be attainable with relatively small control currents for magnetic levitation systems.

Since the preceding analysis was based on a linear approximation of the dynamic equations of the levitating coil and on a linear analysis of the control, it is of interest to determine the transient response of the non-linear equations of motion (3.7) with the linear feedback control on the current. This was done by numerically integrating the equations

$$\dot{x}_1(t) = x_2(t) \tag{4.12}$$

$$\dot{x}_2(t) = -\frac{(Ie + \mathscr{k} x_2(t))^2}{m} \cdot \eta \cdot \frac{dM}{d\hat{z}} - g$$

for various values of $\mathscr{k}$ (i.e., damping) and various initial vertical position errors. The response to an initial height error of 1 inch without control and with $\mathscr{k}$ (or β) adjusted to drive disturbances to zero in 5 sec. are shown in Figure 3. Because the undamped oscillation is not symmetric about the zero axis it is apparent that the response of the non-linear equations differs from that predicted by the linear analysis. However, the frequency is the same as that determined from the linear analysis. The maximum control current used to attain the damped response was 1.4 amperes which is 2.14% of the equilibrium current. This peak value is slightly less than the 2.28% predicted by the linear analysis. When the coil was subjected to an initial height error of -1 inch, a similar type response to that in Figure 3 was obtained but the maximum control current was 1.8 amperes or 3.64% of the equilibrium current. It is obvious from both these results and physical reasoning that the linear approximation is less accurate for displacements toward the conducting plate than for those away from it. This would follow from the fact that the force law is more non-linear the smaller z(t) is, i.e., the closer the coil is to the plate, and more energy is added to the system by moving the coil toward the plate than away from it.

The response of the non-linear equations (4.12) with $\mathscr{k}$ adjusted to drive disturbances to zero in 1 sec. were also obtained. In this case, the maximum value of the control current was found to be 8.45% of the equilibrium current. Thus, the peak value of 7.3% predicted by the linear analysis is not exact but is quite accurate.

In summary, the feedback gain needed to damp out a disturbance in a given time can be accurately determined by Eqs. (4.6) and (4.10). A reasonable approximation to the maximum value of the resulting control current is then given by (4.11).

V. EXPERIMENTAL STUDY OF CONTROL

In order to assess the accuracy of the theoretical calculations of the control currents needed for vertical damping, an experimental feedback control demonstration was built. A 175 turn coil was wound around the perimeter of a 7/8 inch x 1-7/8 inch x 3/8 inch ceramic magnet. This magnet system was attached to one end of a lever arm and was placed over a spinning 6 inch radius aluminum disk. The permanent magnet provided the levitation force and the coil carried the feedback current needed for damping. The signal proportional to velocity which is needed for the feedback control was obtained by amplifying the voltage induced across a small coil suspended from the opposite end of the lever arm from the levitated magnet and which moved up and down in the field inside two ceramic ring magnets. This voltage was amplified by two operational amplifier stages and one power amplifier stage and was placed across the control coil. The disk was rotated at 4,400 rpm which corresponds to a speed at its outer edge of 157 mph. The equilibrium height of the magnet above the spinning disk was about 1 inch. The vertical motion of the levitated magnet when it was perturned was very oscillatory with a frequency of 2.6 hertz. While the frequency of oscillation is theoretically independent of the coil mass, it does depend on the geometry of the coil and the height. Thus it is to be expected that the frequency in the experimental setup would differ from that in the theoretical analysis due to the difference in the geometry.

Without feedback control of the coil current, oscillations resulting from perturbations took of the order of a few hundred cycles to damp out. With the feedback control, oscillations were damped out in less than 1 cycle. The peak control current needed to attain this damping was .576 amperes. The equilibrium current (in a current sheet around the perimeter) for a ceramic magnet of the size indicated is approximately 2,000 amperes. Thus, the number of ampere turns used for the feedback control was 5.0% of the equilibrium ampere turns. This compares with the 6.4% predicted by the theoretical analysis of the linearized equations in the previous section to damp out oscillations in about 2 cycles. The agreement between these experimental results and the theoretical predictions is very good. It confirms that excellent damping of the oscillatory vertical motion of levitated magnets can be achieved by controlling relatively small currents in the levitation coils (or perhaps in trim coils near the levitation coils).

VI. AN ANALYSIS OF SEPARATE CONTROL COILS

The analyses of the previous sections assumed that the current in a levitation coil can be varied by feedback control in order to obtain sufficient damping of the vertical motion of magnetically levitated vehicles. However, it is evident from the equilibrium levitation currents required for a reasonably sized vehicle (e.g., 100,000 lb.) that the power per levitation coil on such a vehicle would be of the order of megawatts if ordinary copper wire at ambient temperatures was used for the coil. Thus, it is almost imperative to consider the case where superconducting coils at liquid Helium temperatures are used for levitation to minimize the power requirement for levitation. However, certain practical difficulties may arise in trying to control the current in superconducting coils according to a feedback control law rather than operating the coils in the persistent mode. Thus, it is important to investigate the possibility of using separate control coils at ambient temperature to achieve the desired vertical damping of a levitated vehicle for which the equilibrium levitation force is obtained using a superconducting coil operating in the persistent mode. In this section, an initial analysis of the feasibility of such an approach is made.

The configuration of levitation and control coils to be analyzed is shown in Figure 4. The following assumptions are made in the analysis:

(i) the levitation and control coils are rigidly connected.

(ii) vertical motion only is considered

(iii) there is no common flux linkage between levitation and control coils

(iv) levitation and control coils have the same dimensions and number of turns

(v) there are at least two levitation/control coil sets per vehicle.

Because of assumptions (i) and (v) rotations of the control coil/levitation coil combination can essentially be ignored and assumption (ii) is valid. Assumption (iii) must be made since any common flux linkage between the control and levitation coils would make the analysis more difficult. This is because a superconducting coil in the persistent mode will tend to conserve flux linkages through it and the current in the levitation coil would adjust to offset any changes in flux linkage through it tending to be caused by the control coil. Finally assumption (iv) is optimistic in the sense that the control coil would be less effective if it was smaller and had fewer turns than the levitation coil. Thus, the results of this analysis should be considered as an upper bound on the performance of separate control coils.

The force in the z direction for the two magnet system in Figure 4 subject to the assumptions stated above can be written as

$$F_z = - I_e{}^2 \frac{dM_{\ell\ell'}}{d\hat{z}} - 2I_e i(t) \frac{dM_{\ell c'}}{d\hat{z}} - i^2(t) \frac{dM_{cc'}}{d\hat{z}} - mg, \tag{6.1}$$

where $M_{\ell\ell'}$ = mutual inductance between the levitation coil and its image, $M_{\ell c'}$ = mutual inductance between the levitation coil and the image of the control coil, $M_{cc'}$ = mutual inductance between the control coil and its image. The expressions for these mutual inductances as a function of the separation of the coils can be obtained using a more general expression than (3.1) which is given in [5]. I_e is the equilibrium current chosen so that for i(t) nominally zero, I_e levitates the coil at height h. Thus,

$$I_e{}^2 = - \frac{mg}{\left.\frac{dM_{\ell\ell'}}{d\hat{z}}\right|_{2h}} \tag{6.2}$$

as in the previous analyses. If the same coordinates as used in (3.4) - (3.6) are used, the non-linear state equations describing the vertical motion of the magnets are

$$\dot{x}_1(t) - x_2(t) \tag{6.3a}$$

$$\dot{x}_2(t) = - \eta \frac{I_e^2}{m} \frac{dM_{\ell\ell'}}{d\hat{z}} - 2\eta \frac{I_e i(t)}{m} \frac{dM_{\ell c'}}{d\hat{z}} - \eta \frac{i^2(t)}{m} \frac{dM_{cc'}}{d\hat{z}} - g. \tag{6.3b}$$

If $x_1(t)$ and $x_2(t)$ are assumed to be small deviations from the desired steady motion, $z(t) = h$ and $\dot{z}(t) = 0$, and $i(t)$ is assumed to be a relatively small control current equal to $\hat{k}x_2(t)$, then the equations of motion (6.3a) and (6.3b) can be approximately represented by the linear equations

$$\dot{x}_1 = x_2 \tag{6.4a}$$

$$\dot{x}_2(t) = - \alpha x_1(t) - \hat{\beta} x_2(t) \tag{6.4b}$$

where

$$\alpha = \frac{2I_e^2 \eta}{m} \left. \frac{d^2 M_{\ell\ell'}}{d\hat{z}^2} \right|_{\hat{z} = 2h}$$

and

$$\hat{\beta} = \frac{2\eta I_e \hat{k}}{m} \left. \frac{dM_{\ell c'}}{d\hat{z}} \right|_{\hat{z} = 2h} .$$

It is apparent that α is exactly the same as that given in the analyses in the previous sections if it is assumed that the mass of the added control coil is much less than the mass supported by the levitation coil. However, the expression for $\hat{\beta}$ differs from that for β.

It turns out that a linear analysis of the control currents as was made in Section IV is not very accurate for determining the peak value of the control current in the separate control coil. This is because the term $\frac{\eta i^2(t)}{m} \frac{dM_{cc'}}{d\hat{z}}$, which is zero to a first order approximation near equilibrium, is in fact significant with respect to the other terms in (6.3b). Thus, while $\hat{\beta}$ can be defined as was β in (4.6) to achieve a given damping time, the resulting control current is much less than that predicted by a linear analysis. The maximum current must be determined by solution of the non-linear equation (6.3). The solutions to (6.3) for an initial height error of 1 inch and for various values of the feedback gain were obtained numerically. The control current required in the levitation coil and in a separate control coil for a damping time of 5 sec. are shown in Fig. 5. This control coil was situated directly beside the levitation coil. The phase difference is due to the sign of $\frac{dM_{\ell c'}}{d\hat{z}}$ in the expression for $\hat{\beta}$. It is seen that the maximum control current needed in the separate coil is 4.28 times that which would be needed in the levitation coil itself (or 9.25% of the equilibrium current). For the case of damping in 1 sec. the ratio was found to be 3.02. Thus, while the control currents required in separate control coils are considerably higher than those that would be required in the levitation coils to achieve the same damping, they may not be prohibitive. The final design for a magnetic suspension system will depend on the tradeoff between the higher control power required by separate control coils and the advantage gained by having the main levitation coils operate in the persistent mode.

VII. CONCLUSIONS

From the analyses summarized in this paper, it can be concluded that excellent damping of the vertical motion of levitated magnets can be achieved by implementing a feedback control of the current in the levitation coil. The variations in the coil current needed for control were predicted to be of the order of 2% to 5% of the current required for levitation, and this agreed very well with the experimental results obtained using a small model of a levitation and control scheme. However, it was shown that the currents required for control are greater if control coils separate from the levitation coils are used. Since the levitation coils would probably have to be superconducting coils at liquid Helium temperature, the practical problems of implementing a feedback control on the levitation coil current must be weighed against the greater power required by separate control coils. Because of the relatively small degree of coupling between the horizontal and vertical motion of levitated coils and because the feedback control tends to decrease this coupling further, it appears that the lateral, longitudinal, and vertical motion of levitated vehicles can be analyzed and controlled independently.

Further work is required with respect to the problems of lateral guidance, propulsion, failure modes, etc., associated with magnetically levitated vehicles before the feasibility of such systems is firmly established.

ACKNOWLEDGMENT

The author would like to acknowledge many helpful discussions with Drs. L. C. Davis, J. R. Reitz and R. H. Borcherts.

REFERENCES

1. Survey of Technology for High Speed Ground Transport, Pt. 1, U. S. Department of Commerce, Washington, D. C., June, 1965, p. v-180.

2. J. R. Powell and G. R. Danby, "High Speed Transport by Magnetically Suspended Trains", ASME Winter Annual Meeting, New York, Railroad Div., 66-WA/RR-5, Nov. 1966.

3. G. R. Polgreen, "Guided Land Transport", Proc. Institute Mechanical Engineers, Pt. 3G, (London), vol. 181, 1966-67, pp. 145-150.

4. H. T. Coffey, F. Chilton, and T. W. Barbee, Jr., "Suspension and Guidance of Vehicles by Superconducting Magnets", (abs.) J. Appl. Phys., vol. 40, 1969, p. 2161.

5. John R. Reitz, "Forces on Moving Magnets Due to Eddy Currents", J. Appl. Physics, vol. 41, No. 5, April 1970, pp. 2067-2071.

6. Robert H. Borcherts and John R. Reitz, "High Speed Transportation via Magnetically-Supported Vehicles. A Study of Magnetic Forces", to be published in Transportation Research.

7. L. C. Davis and Dennis F. Wilkie, "Analysis of the Motion of Magnetic Levitation Systems: Implications for High Speed Vehicles", submitted to J. Appl. Phys., 1971.

8. Paul B. Bailey and Frederick R. Norwood, "Stability of Magnets Moving above a Conducting Plane", J. Appl. Phys., vol. 41, no. 12, November 1970, pp. 4890-4892.

FIGURE CAPTIONS

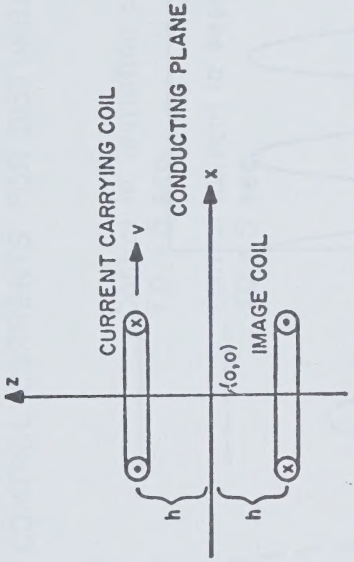
Fig. 1. Coordinate system for a wire moving over an infinite conducting plate.

Fig. 2. Coordinate system for a current carrying coil moving over an infinite conducting plate.

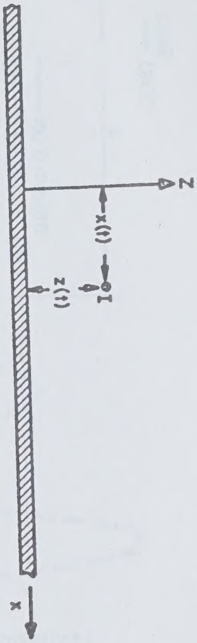
Fig. 3. Motion of a coil sub_ ct to an initial height disturbance with and without feedback control of the current. (2 meter square coil of 4,000 turns supporting 5,000 kg. mass at nominal height of .1m above a plate, nominal velocity 134m/sec.)

Fig. 4. The configuration with separate control and levitation coils.

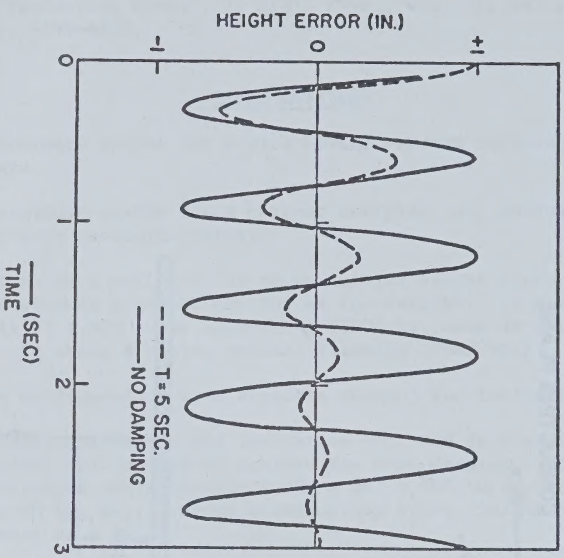
Fig. 5. Control current in the levitation coil and in a separate control coil needed to achieve the same damping. (Control and levitation coils identical, 2m x 2m, 4,000 turns, supporting 50,000 kg. at .1m above a conducting plate, initial disturbances of 1 in.)

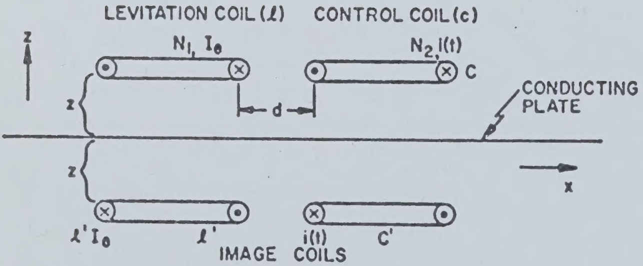


z
CURRENT CARRYING COIL
v
x
CONDUCTING PLANE
x
(0,0)
IMAGE COIL
h
h

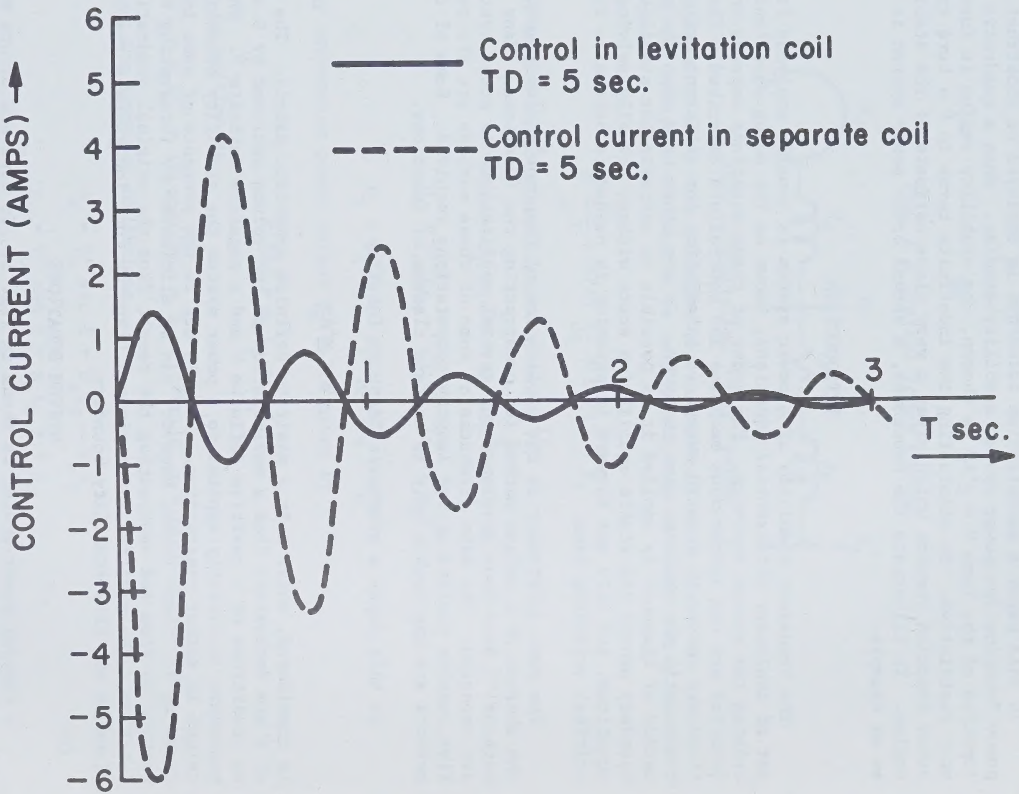


x
z(t)
I
x(t)
Z


HEIGHT ERROR (IN.)
-1
0
+1
0
1
2
3
TIME (SEC)
T=5 SEC.
NO DAMPING


LEVITATION COIL (ℓ)
CONTROL COIL (c)
N1, I0
N2, I(t)
C
z
z
d
CONDUCTING PLATE
z
x
ℓ' I0
ℓ'
i(t)
C'
IMAGE COILS

CONTROL CURRENTS FOR DISTURBANCES



A MAXIMIZED LIAPUNOV FUNCTION FOR POWER SYSTEM STABILITY STUDY

YAO-NAN YU and ALI A. METWALLY
Department of Electrical Engineering,
University of British Columbia,
Vancouver 8, B.C., Canada.

ABSTRACT

In this paper a maximization technique is employed to construct a Liapunov function for power system stability studies. When a quadratic Liapunov function of the form $V = \underline{x}'A\underline{x}$ is chosen, the stability region is found to be very restrictive. By eliminating the indefinite terms in $\dot{V}$ a Luré type Liapunov function results which gives a very close estimate of the stability region. To illustrate the procedure, a second order power system is chosen as an example.

INTRODUCTION

The transient stability of a power system is usually analyzed from a set of nonlinear differential equations, known as the swing equations, describing the system dynamics. The order of these equations depends on how detailed are the synchronous machines and controllers described. The equations are solved by forward sweeps of integration for different switching times until the critical time is found. On the other hand when the second method of Liapunov is applied it is possible to estimate the stability boundary around the stable equilibrium state without actually solving the equations, and only one forward integration is needed to determine the critical switching time.

The main difficulty in the application of Liapunov's direct method is the absence of a unique method for constructing the functions. Many methods[1-7] have been developed and several applications to power systems[8-16] are reported. The main drawbacks of some of these methods are the restrictive results yielded or the lengthy computations required. Some of these methods are applicable only to certain classes of functions.

In this paper a quadratic Liapunov function

$$V = \underline{x}'A\underline{x} \tag{1}$$

is considered, where A is a positive definite symmetric matrix. The elements of A are determined from a maximization of the volume enclosed by V subject to conditions of a positive definite V and a negative definite $\dot{V}$. When this procedure is directly applied to a power system the stability boundary obtained is rather restrictive. This is due to the presence of some indefinite terms in $\dot{V}$. These terms, however, can be eliminated by integrating them with respect to time and subtracting the result from the original quadratic V-function. The modified and maximized Liapunov function gives a result very close to the true stability boundary.

SYSTEM EQUATIONS

A typical power system is shown in Fig. 1. It consists of one machine, with a local load, connected to an infinite system through parallel transmission lines. Fig. 2 is the equivalent of Fig. 1.

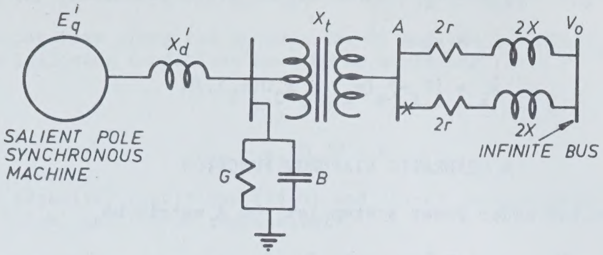


Fig. 1 A Typical Power System

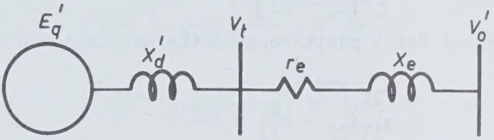


Fig. 2 Equivalent Power System

Assuming that both flux linkages and mechanical power input are constant, and the armature resistance is negligible, the system's swing equation may be written

$$M \frac{d^2\delta}{dt^2} + D(\delta)\frac{d\delta}{dt} + P_e(\delta) = P_i \tag{2}$$

Where $D(\delta)$ is the damping coefficient given by[17],

$$\begin{aligned} D(\delta) &= D_1 \cos^2\delta + D_2 \sin^2\delta \\ D_1 &= V_o'^2(x_q - x_q'')\tau_{qo}''/(x_e + x_q)^2 \\ D_2 &= V_o'^2(x_d' - x_d'')\tau_{do}''/(x_e + x_d')^2 \end{aligned} \tag{3}$$

The energy conversion power output $P_e(\delta)$ is given by

$$P_e(\delta) = v_d i_d + v_q i_q \tag{4}$$

which can be written as

$$P_e(\delta) = B_1 E_q'^2 + [B_2\cos(\delta+\beta) + B_3\sin(\delta+\gamma)]E_q' + B_4\sin(\delta+\gamma)\cos(\delta+\beta) \tag{5}$$

by applying the terminal conditions

$$\begin{aligned} v_d &= V_o' \sin\delta + r_e i_d - x_e i_q \\ v_q &= V_o' \cos\delta + x_e i_d + r_e i_q \end{aligned} \tag{6}$$

and the synchronous machine phasor relations

$$\begin{aligned} v_d &= x_q i_q \\ v_q &= E_q' - x_d' i_d \end{aligned} \tag{7}$$

Choosing $x_1 = \delta-\delta_o$ and $x_2 = d\delta/dt$ as the state variables, (2) becomes

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= [P_i-P_e(x_1) - x_2D(x_1)]/M \end{aligned} \tag{8}$$

A QUADRATIC LIAPUNOV FUNCTION

For the second order power system let the A matrix be

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \tag{9}$$

The Sylvester conditions for a positive definite matrix A are

$$a_{11} > 0 \tag{10-a}$$

$$a_{11}a_{22} - a_{12}^2 > 0 \tag{10-b}$$

The time derivative of (1) for the second order system (8) is given by

$$\begin{aligned} \dot{V} &= 2[(a_{11}x_1+a_{12}x_2)\dot{x}_1 + (a_{12}x_1+a_{22}x_2)\dot{x}_2] \\ &= 2\{[a_{12}-a_{22}D(x_1)/M]x_2^2 +[(a_{11}-a_{12}D(x_1)/M)x_1 + a_{22}(P_i-P_e(x_1))/M]x_2 \\ &\qquad + a_{12}x_1[P_i - P_e(x_1)]/M\} \end{aligned} \tag{11}$$

To ensure that $\dot{V}$ is negative definite in a region around the origin it is required that $\dot{V}_2$, the second degree terms of the Taylor series of $\dot{V}$, be negative definite. The Taylor series expansion of (8) is

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = \sum_{i=1}^{\infty} p_i x_1^i + x_2 \sum_{i=1}^{\infty} q_i x_1^{i-1}$$

$$\begin{aligned} q_1 &= -[(D_1+D_2) + (D_1-D_2)\cos 2\delta_o]/2M \\ q_i &= -[2^{i-2}(D_1-D_2)\cos(2\delta_o+ \tfrac{i-1}{2}\pi)]/M(i-1)!, \ i=2,3,\ldots \\ p_i &= \{E'_q[B_3\cos(\delta_o+\gamma+ \tfrac{i-1}{2}\pi)-B_2\sin(\delta+\beta+ \tfrac{i-1}{2}\pi)] \\ &\qquad +2^{i-1}B_4\cos(2\delta_o+\beta+\gamma+ \tfrac{i-1}{2}\pi)\}/iM , \ i=1,2,\ldots \end{aligned} \tag{12}$$

and $\dot{V}_2$ has the form

$$\dot{V}_2 = 2\underline{x}' \begin{bmatrix} a_{12}p_1 & \frac{1}{2}(a_{11}+a_{12}q_1+a_{22}p_1) \\ \frac{1}{2}(a_{11}+a_{12}q_1+a_{22}p_1) & (a_{12}+a_{22}q_1) \end{bmatrix} \underline{x} \tag{13}$$

resulting in the negative definite conditions

$$a_{12}p_1 < 0 \tag{14-a}$$

$$-p_1(a_{11}a_{22}-a_{12}^2) - (a_{11}+a_{12}q_1-a_{22}p_1)^2/4 > 0 \qquad (14\text{-b})$$

For $\dot{V}$ to be negative along the x_1-axis ($x_2=0$) and the x_2-axis($x_1=0$) respectively, the following conditions have to be satisfied

$$a_{12} > 0 \qquad (14\text{-c})$$

$$a_{12} - a_{22}D_2/M < 0 \qquad (14\text{-d})$$

Since p_1 is negative, conditions (14-b) and (14-c) automatically satisfy conditions (10-b) and (14-a) respectively.

The hypervolume enclosed by (1) is found in general as[18]

$$\begin{aligned} I &= 2^n\left(\frac{\pi}{2}\right)^{\frac{n-1}{2}} \frac{1}{1.3.5\ldots n}\sqrt{V^n/|A|} && n \text{ odd} \\ & \ 2^n\left(\frac{\pi}{2}\right)^{n/2} \frac{1}{2.4.6\ldots n}\sqrt{V^n/|A|} && n \text{ even} \end{aligned} \qquad (15)$$

where n is the order of the system. For the second order system, the area enclosed by V is

$$I = \pi V/\sqrt{(a_{11}a_{22}-a_{12}^2)} \qquad (16)$$

This area is then maximized subject to the inequality constraints (10-a), (14-b), (14-c), (14-d) and the tangent conditions

$$\begin{aligned} g_1 &= \dot{V} = 0 \\ g_2 &= \frac{\partial V}{\partial x_1}\cdot\frac{\partial \dot{V}}{\partial x_2} - \frac{\partial V}{\partial x_2}\cdot\frac{\partial \dot{V}}{\partial x_1} = 0 \end{aligned} \qquad (17)$$

The inequality constraints are transformed into equality constraints by introducting the free variables y_i, $i = 1, \ldots, 4$, as follows

$$\begin{aligned} g_3 &= a_{11}-y_1^2 - \varepsilon = 0 \\ g_4 &= a_{12}-y_2^2 - \varepsilon = 0 \\ g_5 &= a_{22}D_2/M-a_{12}-y_3^2- \varepsilon = 0 \\ g_6 &= -p_1(a_{11}a_{22}-a_{12}^2)-(a_{11}+a_{12}q_1-a_{22}p_1)^2/4 - y_4^2 - \varepsilon = 0 \end{aligned} \qquad (18)$$

where ε is a small positive constant.

The gradient projection method[19,20] is employed to maximize (16) subject to constraints (17) and (18) in an augmented space $Z = \{x,a,y\}$.

The numerical example considered has the following data

M	E'_q	P_i	δ_o	β	γ	D_1	D_2	B_1
0.02121	1.053p.u.	0.753p.u.	53.9°	75.35°	78.8°	0.556	0.09	0.244

B_2	B_3	B_4
-0.95	-0.058	0.227

The resulting V-function, $V_m = 86$, is tangent to $\dot{V} = 0$ at

$$x_1 = 0.394 \quad , \qquad x_2 = 1.238$$

and the elements of matrix A are

$$a_{11} = 430 \quad , \quad a_{12} = 1.002 \quad , \quad \text{and} \quad a_{22} = 12.02$$

The resulting stability region is shown in Fig. 3 along with the true stability boundary for comparison.

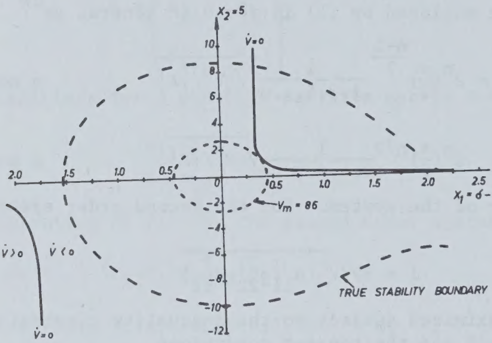


Fig. 3 Stability Boundary of a Quadratic V-Function

A LUR'E TYPE LIAPUNOV FUNCTION

It is noticed that the poor estimation of stability region in Fig. 3 is due to the fact that $\dot{V}$ is negative definite in a small region only, where $\dot{V}_2$ is dominant. This can be overcome by eliminating the indefinite terms in $\dot{V}$ by integrating them with respect to time and subtracting the result from the original V.

The indefinite terms of (11) are

$$\psi = 2\{[a_{11}-a_{12}D(x_1)/M]x_1 + a_{22}[P_i-P_e(x_1)]/M\}x_2 \tag{19}$$

The time integration gives

$$\int\psi dt = [a_{11}-a_{12}(D_1+D_2)/2M]x_1^2 - f(x_1) \tag{20}$$

$$f(x_1) = a_{12}(D_1-D_2)[\cos(2x_1+2\delta_o)+2x_1\sin(2x_1+2\delta_o)-\cos(2\delta_o)]/4M$$

$$+ 2a_{22}\{E'_q[B_2(\sin(x_1+\delta_o+\beta)-x_1\cos(\delta_o+\beta)-\sin(\delta_o+\beta))$$

$$-B_3(\cos(x_1+\delta_o+\gamma)+x_1\sin(\delta_o+\gamma)-\cos(\delta_o+\gamma))]$$

$$-B_4[\cos(2x_1+2\delta_o+\beta+\gamma)+2x_1\sin(2\delta_o+\beta+\gamma)-\cos(2\delta_o+\beta+\gamma)]/4\}/M \tag{21}$$

Subtracting (20) from (1) results in

$$V = \underline{x}' \begin{bmatrix} a_{12}(D_1+D_2)/2M & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \underline{x} + f(x_1) \qquad (22)$$

which is a Liapunov function of the Luré type. The time derivative is

$$\dot{V} = 2\{[a_{12}-a_{22}D(x_1)/M]x_2^2 + a_{12}x_1[P_i-P_e(x_1)]/M\} \qquad (23)$$

which is negative definite for

$$-(2\pi-\delta^{us}+\delta_o) < x_1 < (\delta^{us}-\delta_o) \qquad (24)$$

The numerical maximum of V, V_{max}, that describes a closed curve, is determined as follows. Equation (22) is rearranged to read,

$$x_2 = \{-a_{12}x_1 \pm \sqrt{[a_{12}^2-a_{12}a_{22}(D_1+D_2)/2M]x_1^2-a_{22}f(x_1)+a_{22}V}\}/a_{22} \qquad (25)$$

For this equation to describe a closed curve, for each value of V, there must be a value of x_1 such that the square root is equal to zero, which gives

$$V = f(x_1) + [a_{12}(D_1+D_2)/2M - a_{12}^2/a_{22}]x_1^2 \qquad (26)$$

By differentiating (26) with respect to x_1 and equating to zero, we have

$$x_1 = (a_{22}^2/a_{12})[P_i-P_e(x_1)]/[a_{22}D(x_1) - a_{12}M] \qquad (27)$$

V_{max} is found by solving (27) for x_1 and substituting the result in (26). This value of V is not necessarily the one that bounds the stability region since it may intersect $\dot{V} = 0$. Therefore the value of V tangent to $\dot{V} = 0$, V_T, is obtained by solving (17) for x_1 and x_2 and then substituting into (22). Finally the value of V describing the stability boundary, V_m, is determined from V_{max} and V_T whichever is smaller.

Following the new procedure for the same power system investigated in the preceding section results in $V_m = 878.4$, which is tangent to $\dot{V} = 0$ at

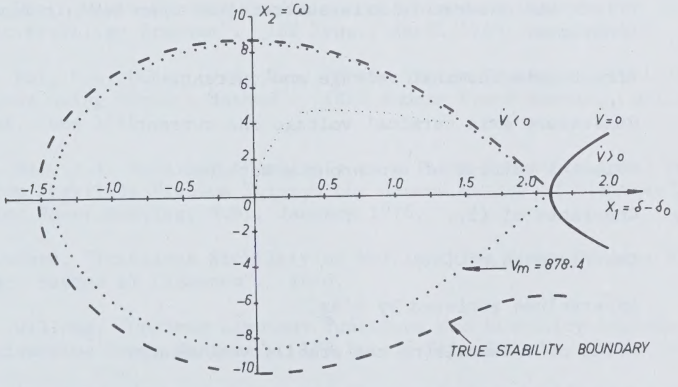


Fig. 4 Stability Boundary of Modified V-Function

$$x_1 = 2.1 \quad , \quad x_2 = 0$$

and the elements of matrix A are the same except for a_{11} which is given by $a_{12}(D_1+D_2)/2M$. Fig. 4 shows the resulting stability region to be very close to the true stability boundary.

CONCLUSION

A new procedure for constructing a Liapunov function V for power systems is developed. It starts with a quadratic form and is modified by eliminating the indefinite terms of $\dot{V}$ to ensure its negative definiteness. The resulting Liapunov function is of the Luré type. The hypervolume of V is then maximized. Applying this procedure to a single-machine power system it is found that the resulting region of stability is very close to the true boundary. The method can be readily extended to multimachine power systems[21]. It is also found that a quadratic V-function by itself does not yield a good estimate of the stability region for a power system.

NOMENCLATURE

δ	Angle between the quadrature axis of synchronous machine and infinite bus voltage.
δ_o, δ^{us}	Values of δ at stable and unstable equilibria respectively.
M	Inertia constant.
P_i	Mechanical power input.
V'_o	Infinite bus voltage.
x'_d, x''_d	Direct axis transient and subtransient reactances.
x_q, x''_q	Quadrature axis transient and substransient reactances.
x_e, r_e	External system equivalent reactance and resistance.
τ''_{do}, τ''_{qo}	Direct and quadrature axis subtransient open circuit time constants.
V_d, i_d	Direct axis terminal voltage and current.
V_q, i_q	Quadrature axis terminal voltage and current.
E'_q	Internal voltage of synchronous machine.
B_1-B_4, β, γ	Constants of (5).
p_i,q_i	Coefficients of (12).
I	Hypervolume enclosed by $\underline{x}'A\underline{x}$.
V_m	Value of V describing the stability boundary.
V_{max}	Numerical maximum of V describing a closed curve.
V_T	Value of V tangent to $\dot{V} = 0$

REFERENCES

1. M.L. Cartwright, "On the Stability of Solutions of Certain Differential Equations of Fourth Order". Quart. J. Mech. Appl. Math., 1956, 9, (2).

2. D.R. Ingwerson, "A Modified Liapunov Method for Nonlinear Stability Analysis", IRE Trans. on Automatic Control, May 1961, pp. 199-210.

3. D.G. Schultz, J.E. Gibson, "The Variable Gradient Method for Generating Liapunov Functions", AIEE Trans. on Automatic Control, Sept. 1962.

4. G.P. Szego, "A Contribution to Liapunov's Second Method: Nonlinear Autonomous Systems", J. of Basic Engineering, Dec. 1962, pp. 571-578.

5. S.G. Margolis, W.G. Vogt, "Control Engineering Applications of V.I. Zubov's Construction Procedure for Liapunov Functions", IEEE Trans. on Automatic Control, April 1963, pp. 104-113.

6. R.E. Kalman, "Liapunov Functions for the Problem of Luré in Automatic Control" Proc. Nat. Acad. Sci., U.S. 49, 2, 1963, pp. 201-205.

7. C.S. Chen, E. Kinnen, "Construction of Liapunov Functions", J. of Franklin Inst., Vol. 289, No. 2, Feb. 1970, pp. 133-146.

8. A.H. El-Abiad, K. Nagappan, "Transient Stability Regions of Multimachine Power Systems", IEEE Trans. on Power Apparatus and Systems, Vol. PAS-85, No. 2, Feb. 1966, pp. 169-179.

9. G.E. Gless, "Direct Method of Liapunov Applied to Transient Power System Stability", ibid. pp. 158-168.

10. Y.N. Yu, K. Vongsuriya, "Nonlinear Power System Stability Study by Liapunov Function and Zubov's Method". IEEE Trans. on Power Apparatus and Systems, Vol. PAS-86, No. 12, Dec. 1967, pp. 1480-1485.

11. M.W. Siddiquee, "Transient Stability of an A.C. Generator by Liapunov's Direct Method." Int. J. of Control, Vol. 8, No. 2, 1968, pp. 131-144.

12. N.D. Rao, "Routh-Hurwitz Conditions and Liapunov Methods for the Transient Stability Problem". IEE Proc., April 1969, pp. 537-547.

13. M.A. Pai, M.A. Mohan, J.G. Rao, "Power System Transient Stability Regions using Popov's Method". IEEE Summer Power Meeting, Dallas, Texas, June 1969.

14. N.D. Rao, A.K. DeSarkar, "Analysis of a Third Order Nonlinear Power System Stability Problem Through the Second Method of Liapunov". IEEE Winter Power Meeting, N.Y., January 1970.

15. G. Luders, "Transient Stability of Multimachine Power Systems Via the Direct Method of Liapunov". ibid.

16. J.L. Willems, "Optimum Liapunov Functions and Stability Regions for Multimachine Power Systems". IEE Proc., Vol. 117, No. 3, March 1970.

17. R.V. Shepherd, "Synchronizing and Damping Torque Coefficients of Synchronous Machines". AIEE Trans. on Power Apparatus and Systems, Vol. 80, June 1961, pp. 180-189.

18. Ali A. Metwally, "Power System Stability Study by Szego's Method and a Maximized Liapunov Function". M.A.Sc. Thesis University of British Columbia, December 1970.

19. J.B. Rosen, "The Gradient Projection Method for Nonlinear Programming. Part I: Linear Constraints". J. Soc. Industrial and Applied Math., No. 8, 1960, pp. 181-217.

20. J.B. Rosen, "The Gradient Projection Method for Nonlinear Programming, Part II: Nonlinear Constraints". J. SIAM, Dec. 1961, pp. 514-532.

21. Ali A. Metwally, Y.N. Yu, "Optimized Luré Type Liapunov Function for Power System Stability Studies". Paper Submitted to the IEEE Power Engineering Society Winter Meeting, New York, January 1972.

A COMPUTER-AIDED REGULATOR DESIGN*

J. S. Karmarkar
NRC Postdoctoral Research Associate
Ames Research Center (NASA)
Moffett Field, California

D. D. Šiljak
Electrical Engineering Department
University of Santa Clara
Santa Clara, California

ABSTRACT

This paper presents a unified computer-aided method for the design of linear and nonlinear regulators. Starting with the linear vector equations

$$\dot{\underset{\sim}{x}}(t) = \underset{\sim\sim}{P}\underset{\sim}{x}(t) + \underset{\sim}{q}u(t)\ ,\ u(t) = -\underset{\sim}{r}^T\underset{\sim}{x}(t)$$

an algorithmic procedure is developed for selecting the regulator parameter vector $\underset{\sim}{r}$, within the framework of classical time and frequency domain specifications, as well as the quadratic cost optimality criterion.

For nonlinear regulators described by nonlinear vector equations

$$\dot{\underset{\sim}{x}}(t) = \underset{\sim\sim}{P}\underset{\sim}{x}(t) + \underset{\sim}{q}\phi[u(t)]\ ,\ u(t) = -\underset{\sim}{r}^T\underset{\sim}{x}(t)\ ,$$

attention is focused on the design of absolutely stable Lur'e regulators subject to prescribed exponential stability and sector maximization requirements.

Reformulating the performance specifications, in terms of a set of inequalities, delineates a feasible region in the regulator parameter vector space. Then, maximizing the volume of an imbedded hypercube (or hypersphere) inside the region, via mathematical programming methods, results in an easily visualized solution set; this feature is particularly attractive to the practical engineer, aware of real-world implementation tolerances and system parameter uncertainities.

I. INTRODUCTION

In order to realistically design, increasingly complex systems, arising from modern technology, subject to a myriad of performance specifications, it has become imperative that the computer play a central role in any successful design algorithm. This approach to system design is in the spirit of the newly emerging and already fruitful field of computer-aided circuit design†. At this stage, however, it suffers from the same drawback of being wordy, where concise mathematical statements would otherwise suffice.

Partitioning systems into linear and essentially nonlinear, reinterpreting known results in control theory [1-4], a comprehensive design algorithm is developed. More precisely, the time and frequency domain specifications, e.g., exponential stability, damping factor, steady state error, frequency response peak, time response peak, dominant root location and sensitivity,

---

[*] This work was supported by NASA Grant NGR 05-017-010.

[†] IEEE Transactions on Circuit Theory, edited by R. Rohrer, Vol. CT-18, No. 1, January, 1971.

mean squared error, integral squared error; as well as the quadratic index optimality criterion (stated in terms of the return difference concept), for linear systems, are formulated in a mathematical programming format [5]. Moreover, for nonlinear Lur'e class systems the reformulation is accomplished using the absolute stability results subject to exponential stability and sector maximization requirements [1-2].

Essentially, the preceding performance specifications yield a set of inequalities, delineating a feasible region in the space of the design variables called the regulator parameter vector space. Consequently maximizing the volume of a hypercube (linear/nonlinear inequalities) or hypersphere (linear inequalities) imbedded in this feasible region completes the design - the specifications are satisfied for all regulator vectors lying within the imbedded volume.

It cannot be overemphasized that additional constraints on the parameters, arising for reasons of economy, realizability or availability, can be introduced into the design problem formulation in a natural manner, resulting in a more realistic design. Furthermore, although it is sometimes possible to obtain a specific set of numerical values for the parameters, to satisfy design requirements; the practical designer is aware that neither are the design specifications precise, nor is the data regarding the physical system precise; thus he is biased towards methods which provide a solution-set rather than a single point.

Although the main thrust of the present paper is directed toward single input-output systems, the proposed algorithm can be readily extended to multi input-output systems, utilizing the necessary results developed elsewhere [6].

II. IMBEDDING PROBLEM AND SOLUTION

The general mathematical programming problem [7] is to determine a vector $\hat{\underset{\sim}{r}}^* = (r_1^*, \ldots, r_n^*)^T$ that solves the problem of minimizing the objective function $\hat{\underset{\sim}{f}}(r)$ subject to constraints; that is,

$$\text{minimize:} \quad f(\hat{\underset{\sim}{r}}) \tag{1a}$$

$$\text{subject to:} \quad h_i(\hat{\underset{\sim}{r}}) \geq 0, \; i = 1, 2, \ldots, \ell \, . \tag{1b}$$

$$g_j(\hat{\underset{\sim}{r}}) = 0, \; j = 1, 2, \ldots, m \, . \tag{1c}$$

where $\hat{\underset{\sim}{r}} \in R^n$ which is isomorphic to the Euclidean n-space; f, $\{h_i\}$, $\{g_j\}$ are continuous functions.

Consider the problem of imbedding a hypersphere in the feasible region $\mathcal{R}$, delineated by linear inequality constraints of the form,

$$a_{i0} + \sum_{j=1}^{n-1} a_{ij} \, r_j \geq 0 \, , \; i = 1, 2, \ldots, \ell \, . \tag{2}$$

where a_{i0} and a_{ij} are constants. Normalizing the inequalities and introducing an auxiliary variable r_n (to represent the hypersphere radius) using the Chebyshev minimax formulation, the following mathematical programming problem [8] results:

$$\begin{aligned} &\text{minimize:} \quad (-r_n) \\ &\text{subject to:} \left\{ \left(a_{i0} + \sum_{j=1}^{n-1} a_{ij} r_j \right) \Big/ \left(\sum_{j=1}^{n-1} a_{ij}^{\,2} \right)^{1/2} \right\} - r_n \geq 0, \; i=1,2,\ldots,\ell \end{aligned} \tag{3}$$

And (3) can be solved, using the well-known linear programming algorithm [8], yielding $\hat{\underset{\sim}{r}}^* = (\underset{\sim}{r}^*, r_n^*)^T$ where r_n^* represents the maximized hypersphere radius and $\underset{\sim}{r}^* = (r_1^*, r_2^*, \dots, r_{n-1}^*)^T$ the hypersphere center, i.e., the solution set is:

$$\mathcal{R}^* \equiv \{\underset{\sim}{r} : \| \underset{\sim}{r} - \underset{\sim}{r}^* \| < r_n^*\}. \tag{4}$$

Moreover, the side of an inscribed hypercube is $2[r_n^*/(n-1)^{1/2}]$.

It is noteworthy that if $r_n^* \leq 0$, then one concludes that no feasible region $\mathcal{R}$ exists. In section IV, it will be shown that the design of absolutely stable systems with a state variable feedback regulator can be cast into the format of Equation (3).

Now, to generalize the above result to imbedding a hyperellipse of prescribed axial ratios and orientation, a transformation is called for such that (a) hyperellipses are transformed to hyperspheres of equal volume, (b) convexity of the transformed feasible region is retained. It suffices to state that this can be accomplished by applying an appropriate equiaffine transformation [15], to the original set of variables [5].

Proceeding to consider the case occurring most often in practice, namely, the inequalities are nonlinear. It is relevant to determine whether a feasible region exists. This is accomplished by solving the Chebyshev-like problem:

$$\begin{aligned} &\text{minimize:} && (-r_n) \\ &\text{subject to:} && h_i(\underset{\sim}{r}) - r_n \geq 0 \text{ , } i = 1, 2, \dots, \ell . \end{aligned} \tag{5}$$

where $\underset{\sim}{r} = (r_1, \dots, r_{n-1})^T$; and the problem is infeasible if $r_n^* \leq 0$.

Provided the problem is feasible, a hypercube is imbedded in $\mathcal{R}$, using the formulation:

$$\begin{aligned} &\text{minimize:} && (-r_n) \\ &\text{subject to:} && h_1(r_1 - ir_n, r_2 - jr_n, \dots, r_{n-1} - kr_n) \geq 0 \\ & && \vdots \\ & && h_1(r_1 - ir_n, r_2 - jr_n, \dots, r_{n-1} - kr_n) \geq 0 \end{aligned} \tag{6}$$

where i, j, ... , k take on all possible combinations among 1, 0, -1 . And the corresponding solution set is:

$$\mathcal{R}^* \equiv \{\underset{\sim}{r} : | r_i - r_i^* | < r_n^* \text{ , } i = 1, 2, \dots, (n-1)\cdot\} \tag{7}$$

Note that the original ℓ inequalities are each decomposed into $3^{(n-1)}$ inequalities in (6). Although (6) appears formidable, it is easily programmed by implementing (n-1) nested DO loops for each inequality as shown in [5], using the SUMT [7] computer program.

It is emphasized that the above formulation assumes that the region $\mathcal{R}$ delineated by the nonlinear inequalities is convex and therefore to imbed a hypercube, only the vertices of the hypercube need be tested. In practice, neither is it a trivial matter to test for convexity of a given nonlinear function, nor are the inequalities, arising from the simplest practical problems, convex. Consequently, an expedient strategy, employed in (6), is to check additional points lying on the hypercube surface and assume that if the inequality is not violated for these points, then it will not be violated in the interior of the hypercube. Moreover, parallelepiped of prescribed orientation and axial ratios can be imbedded, using the appropriate equiaffine transformation, described earlier.

III. LINEAR SYSTEM DESIGN

Consider a linear single input-output time-invariant system described by the vector differential equation

$$\dot{\underset{\sim}{x}}(t) = \underset{\sim}{P}\,\underset{\sim}{x}(t) + \underset{\sim}{q}\,u(t) \tag{8a}$$

$$u(t) = -\underset{\sim}{r}^T\,\underset{\sim}{x}(t) \tag{8b}$$

where (8a) represents the controlled object, and (8b) the regulator. $\underset{\sim}{P}$ is a constant NxN matrix, $\underset{\sim}{q}$ and $\underset{\sim}{r}$ are constant N-vectors. And $\underset{\sim}{x}(t)$ is a real N-vector representing the state of the system. The corresponding open and closed loop characteristic equations are:

$$\Lambda(\lambda) \equiv \det\,(\lambda\underset{\sim}{I} - \underset{\sim}{P}) = \sum_{k=0}^{N} b_k(\underset{\sim}{r})\,\lambda^k \tag{9}$$

$$\Gamma(\lambda) \equiv \det\,(\lambda\underset{\sim}{I} - \underset{\sim}{P} + \underset{\sim}{q}\underset{\sim}{r}^T) = \sum_{k=0}^{N} c_k(\underset{\sim}{r})\,\lambda^k \tag{10}$$

where $\lambda = -\delta + i\omega$ represents the complex variable. It is well-known that the time and frequency domain performance of a linear system is essentially determined by the zeros of its characteristic polynomial (10). Our concern is to outline the manner in which some of the classical performance criteria can be expressed in terms of algebraic inequalities involving the regulator vector $\underset{\sim}{r}$.

Given (10), exponential stability and damping factor constraints can be expressed as a set of algebraic determinental inequalities, as in [1,5,9]. Moreover, algebraic inequalities for dominant mode design, given the dominant pair locations and sensitivity can be developed using root-coefficient relations [5]. Furthermore, using the final value theorem of complex variable theory, a bound for the output steady state error, given a prespecified deterministic input can be formulated [5]. Finally, frequency and time response peak bounds can also be developed [5].

The classical design problems can now be stated:

Problem A - (Stability Design): Given (10), find the largest hypercube in regulator space, such that prescribed exponential stability/damping factor/steady state error specifications are satisfied; this entails formulating and solving the associated $n = (N+1)$ - dimensional mathematical programming Problem (6).

Problem B - (Dominant Mode Design): Given (10), find the largest hypercube in regulator space, such that prescribed dominant root location and sensitivity specifications are met. In addition, the performance specifications of Problem A and time/frequency domain peak bounds can be superimposed [5]. Again this entails formulating and solving the associated $n = (N+1)$ - dimensional mathematical programming programming problem (6).

Enumerating the several advantages that accrue from the solution set format (7): Any $\underset{\sim}{r}$ satisfying (7) is acceptable; if $\underset{\sim}{r}^*$ is the operating point all parameter variations within (7) are acceptable; r_n^* represents, in quantative terms, the "stringency of the constraints" and nonfeasibility is indicated by $r_n^* \leq 0$. Moreover, note that as the dominancy constraint of Problem B is made increasingly stringent, the solution set (7) will be of decreasing volume; resulting in a trade-off between dominancy and allowable parameter variations. Finally, additional constraints, arising for any reason whatsoever can be introduced naturally into the problem formulation.

Besides the aforementioned classical performance specifications, it is advantageous to ensure that the system (8) is optimal with respect to some quadratic performance of the form [2,10]:

$$J = \int_0^\infty (\underset{\sim}{x}^T \underset{\sim}{Q} \underset{\sim}{x} + u^2)\, dt \tag{11}$$

where $\underset{\sim}{Q} = \underset{\sim}{H}^T \underset{\sim}{H} \geq 0$ is a nonnegative definite matrix. Then, under appropriate controllability and observability conditions [the rational functions $\underset{\sim}{H}^T(\lambda \underset{\sim}{I} - \underset{\sim}{P})^{-1} \underset{\sim}{q}$ have no common cancellable factors], it can be shown [10,11] that the stable control law $u(t)$ of (8b) is optimal with a prescribed degree of exponential stability $\delta \geq 0$, provided

$$\Pi(\omega^2, \underset{\sim}{r}, \delta) \equiv |\Gamma(-\delta+i\omega)|^2 - |\Lambda(-\delta+i\omega)|^2 = \sum_{k=0}^{N-1} a_{2k}(\underset{\sim}{r}, \delta)\omega^{2k} \geq 0 \;, \; \forall \omega \geq 0 \tag{12}$$

Although, given $\underset{\sim}{r}$, a numeric test is available to verify (13); for our purpose it is convenient to use the more conservative result:

A sufficient condition for (8) to be optimal with a prescribed degree of exponential stability $\delta \geq 0$ is that

$$a_{2k}(\underset{\sim}{r}, \delta) \geq 0 \;, \; k = 0, 1, \ldots , (N-1) \tag{13}$$

Clearly, (13) can be introduced into the framework of (6) naturally. Now, it is appropriate to present some justification for designing optimal regulators based on (13). In contrast to the Ricatti equation approach, the conditions on the Q matrix in (11) have been relaxed, in that all possible nonnegative definite $\underset{\sim}{Q}$ matrices are acceptable; this is valuable, since the relation between $\underset{\sim}{Q}$ and the classical performance specifications is, at this time, somewhat obscure. On the other hand, it is known that systems optimal with respect to (11), exhibit several attractive properties. More specifically, from the classical viewpoint, the gain margin is infinite and the phase margin is at least 60°. Furthermore, the system displays tolerance to $(\frac{1}{2}, \infty)$ sector nonlinearities, time delay, and sensitivity reduction with respect to parameter variations [11]. Thus our design philosophy is to retain these attractive features of the quadratic index, without numerically specifying $\underset{\sim}{Q}$ and in addition impose the classical constraints:

Problem C - (Comprehensive Design): Given (8), find the largest hypercube in regulator space, such that prescribed optimality/exponential stability/damping factor/steady state error specifications are satisfied, i.e., solve associated $n = (N+1)$ dimensional mathematical programming problem (6).

Interestingly enough, if all states are not available, frequency domain suboptimality criteria [12] can be cast into this format.

In contrast to the aforementioned optimality constraint which does not consider specific input signals, it is sometimes relevant to require that the integral squared error (ISE) for a deterministic input be less than a prespecified value; this constraint is readily inserted in (6), using standard table of integrals [13]; similarly a squared error (MSE) bound for a statistical input signal with prespecified spectral characteristics can be imposed. Moreover, design situations may arise, wherein it is imperative that, rather than merely place an upper bound, MSE (say) be minimized subject to exponential stability and damping factor requirements [1]; clearly this class of problems is also ameanable to solution in the framework of the general mathematical programming problem (1). Finally, it is emphasized that although the present development used the state representation (8), the present method is applicable to any time-invariant system design, involving static (pure gain) or dynamic (lead/lag filters) compensators.

IV. NONLINEAR SYSTEM DESIGN

Reconsider the system of (8) in the form

$$\dot{x}(t) = P\, x(t) + q\phi[u(t)] \tag{14a}$$

$$u(t) = -r^T\, x(t) \tag{14b}$$

where P is a Hurwitz matrix; the nonlinearity $\phi(u)$ is a real valued, continuous function of a real scalar u, and belongs to the class Φ_κ: $\phi(0) = 0$, $0 \leq u\phi(u) \leq \kappa u^2$, $\kappa < +\infty$.

Provided the solution $x(t,r)$ of (14) is well defined for each r of the parameter space, then system (14) is said to be absolutely stable if the equilibrium $x = 0$ is globally asympototically stable for any $\phi(u) \in \Phi_\kappa$ and $r \in \mathcal{R}$. [1,2].

Assuming (14) to be completely controllable and observable, it can be shown that the system is exponentially absolutely stable provided [1,2,14]:

$$\Pi(\omega^2, r, \kappa, \theta, \delta) \equiv \left|\det(i\omega I - P_\delta)\right|^2 \Big[\kappa^{-1} + \mathrm{Re}\{(1+i\omega\theta)\, r^T(i\omega I - P_\delta)^{-1} q\}$$

$$- \delta\theta\kappa \left| r^T(i\omega I - P_\delta)^{-1} q\right|^2\Big] > 0 \text{ , } \forall\omega \geq 0\text{, holds for some real } \theta \text{ .} \tag{15}$$

And $P_\delta = P + \delta I$ is Hurwitz, where $\delta \geq 0$, is the prescribed degree of exponential stability. Noting that (15) is an even polynomial we have

$$\Pi(\omega^2, r, \kappa, \theta, \delta) \equiv \sum_{k=0}^{N} b_{2k}\, \omega^{2k} \geq 0 \text{ , } \forall\omega \geq 0 \tag{16}$$

where the coefficients $b_{2k} = b_{2k}(r, \kappa, \theta, \delta)$ are real and $b_{2N} \neq 0$. And consequently we obtain the following (somewhat conservative) result [1]:

A sufficient condition for (14) to be exponentially absolutely stable is that $b_{2k}(r, \kappa, \theta, \delta) \geq 0$, $k = 0, 1, \ldots, N$. (17)

It is emphasized that the necessary and sufficient conditions for absolute stability are not known at the present time; thus $\mathcal{R}$ cannot be determined. But an estimate of this region can be obtained using the sufficient conditions (17). The following well-known design formulations result from practical considerations:

Problem D: Given (14), find the regulator vector $r^* \in \mathcal{R}$, such that the sector $[0, \kappa]$ of the class Φ_κ is maximized. Clearly, this can be solved by the nonlinear program, in (N+2) dimension:

$$\text{minimize: } (\kappa^{-1})$$

$$\text{subject to: } b_{2k}(r, \kappa, \delta, \theta) \geq 0 \text{ , } k = 0, 1, \ldots, N \tag{18}$$

Problem E: Given (14) find the largest region $\mathcal{R}^* \subset \mathcal{R}$, such that for all $r \in \mathcal{R}^*$ the system is exponentially absolutely stable, in a specified sector κ_1; the prescribed degree of exponential stability being δ.

It is of particular interest to note that for state variable feedback regulators, i.e., for (14b), the inequalities (17) are linear, provided θ is specified; therefore the hypersphere imbedding scheme and the associated $n = (N+1)$-dimension linear programming formulation (3) is applicable. In general, for each θ_i, we obtain a corresponding $\mathcal{R}^*_{\theta_i}$ and $\mathcal{R}^*$ is the union of these regions, i.e.

$$\mathcal{R}^* = \bigcup \mathcal{R}^*_{\theta_i} \tag{19}$$

where $$\mathcal{R}^*_{\theta_i} = \{\underset{\sim}{r}:\theta = \theta_i\ ,\ \kappa = \kappa_1; ||\underset{\sim}{r} - \underset{\sim}{r}^*|| < r_n^*\} \tag{20}$$

Finally, it is possible to combine Problems D and E to obtain:

Problem F: Given (14), find the largest region $\mathcal{R}^* \subset \mathcal{R}$ such that for all $\underset{\sim}{r} \in \mathcal{R}^*$ and all $\phi(u) \in \Phi_{\kappa_m}$, the maximized class, the system is exponentially absolutely stable.

It cannot be overemphasized that, from a mathematical programming viewpoint, the partitioning of parameters into θ, κ, δ, $\underset{\sim}{r}$ is artificial. Moreover, in practice, the partitioning of (14) into the controlled object and regulator equations is often invalid (for example, when dynamic compensating lead and lag filters are used). Finally, although the regulator equation in this and the preceding sections was restricted to the feedback gain form, in conformity with current literature, the method is by no means restricted to this regulator equation.

IV. COMPUTER ALGORITHM

Keeping in mind the objective of automated design procedures, namely, to free the designer of all tasks, except the qualitative decision making phase; a computer algorithm is outlined in this section. The design cycle, described in the sequel, depicts the manner in which the designer "converges" to an "optimum" design by successive iteration on the design parameters. The various phases of this cycle are considered in some detail.

The control problem formulation phase involves data input to the computer specifying the system configuration, the regulator configuration and the performance specifications to be satisfied.

Based on this information, the computer proceeds to the corresponding programming problem formulation phase, wherein using symbol manipulation algorithms and well-known signal flowgraph methods [16], the set of inequalities resulting from the performance specifications are generated and the imbedding problem is set up.

Subsequently, the computer proceeds to the imbedding problem solution phase, which entails imbedding a hypercube (hypersphere) in the feasible region, using for example, the SUMT [7] program.

Having obtained numerical values for the imbedding problem, simulation of the system for trial points within the hypercube (hypersphere) is automatically initiated; using for example, the 360 CSMP program and the results are displayed on a CRT terminal (ideally) or a plotter.

At this stage, the computer enters the decision phase and program control is transferred to the keyboard, manned by the designer. The designer is now called upon to make a qualitative decision as to whether the current simulation results are satisfactory; if they are not, the designer initiates the modification phase by appropriate alterations in the regulator configuration and performance specifications. His choice being guided by the results of the imbedding, simulation programs and experience.

At the present time, such a comprehensive design package is not available and further work in this direction remains to be done. More specifically, a subprogram for the programming problem formulation phase, using well-known signal flowgraph methods [16], needs to be implemented. Moreover, while the control problem, decision and modification phases, although presently unavailable, do not constitute a major programming effort; the linking of the various phases into a compatible package does not pose difficulties. For the purposes of the present paper, the first two phases were implemented by hand and a straightforward modification of the SUMT [7] program, shown in Figure 1, allowed direct utilization of this program for the imbedding solution phase. It is of interest to note that the SUMT computer code required 80 K bytes of memory and total run time (compile/link/schedule/execute) of the order of 15 seconds on a CDC 6400; while on an IBM 360/67 using double precision, the former was 180 K and the latter approximately 60 seconds.

VI. CONCLUSION

A unified computer-oriented method was proposed for the design of linear time-invariant regulators, for linear and nonlinear systems, using mathematical programming methods.

A unique feature of the formulation was that it allowed one to simultaneously consider classical and modern performance specifications. It was noted that practical constraints arising for reasons of economy, availability and physical realizability could be naturally introduced into the mathematical programming format, resulting in a more realistic design. Moreover, since one obtained a solution-set, rather than a solution-point, practical implementation did not entail accurate parameter settings.

Further work remains to be done in implementing the automated design procedure outlined in Section V and subsequently extending the program to handle multi-input systems, using known results [6].

REFERENCES

[1] Šiljak, D. D., Nonlinear Systems, Wiley, 1969.

[2] Šiljak, D. D., "Algebraic Criterion for Absolute Stability, Optimality and Passivity of Dynamic Systems", Proc. IEE, Vol. 117, No. 10, Oct. 1970, pp. 2033-2036.

[3] Karmarkar, J. S., "Nonlinear Absolutely Stable Regulator Design in Parameter Space", Fourth Hawaii International Conference on System Science, January 12-14, 1971.

[4] Karmarkar, J. S., "Multiparameter Design of Linear Optimal Regulators With a Prescribed Degree of Exponential Stability", Fourth Asilomar Conference on Circuits and Systems, November 18-20, 1970.

[5] Karmarkar, J. S., "A Regulator Design by Mathematical Programming Methods", Ph.D. Dissertation, University of Santa Clara, August 4, 1970.

[6] Šiljak, D. D., "New Algebraic Criteria for Positive Realness", Journal of the Franklin Institute, Vol. 291, No. 2, Feb., 1971, pp. 109-120.

[7] Fiacco, A. V. and McCormick, G. P., Nonlinear Programming: Sequential Unconstrained Minimization Technique, Wiley, 1968.

[8] Zukhovitskiy, S. I. and Aveyeva, L. I., Linear and Convex Programming, Saunders, 1966.

[9] Stojić, M. R. and Šiljak, D. D., "Generation of the Hurwitz, Nyquist and Mikhailov Stability Criteria", IEEE Trans. Auto. Control, AC-10, No. 3, July, 1965, pp. 250-255.

[10] Kalman, R. E., "When is a Linear Control System Optimal?", J. Basic Engrg. Vol. 86, Ser. D, March, 1964, pp. 51-60.

[11] Anderson, B.D.O. and Moore, J. B., "Linear System Optimization with Prescribed Degree of Stability", Proc. IEE, Vol. 116, No. 12, Dec., 1969, pp. 2083-2087.

[12] Canales, R., "A Lower Bound on the Performance of Optimal Regulators", IEEE Trans. Auto Control, AC-15, No. 4, August, 1970, pp. 409-415.

[13] Newton, G. C., Gould, L. A. and Kaiser, J. F., Analytic Design of Linear Feedback Controls, Wiley, 1957.

[14] Šiljak, D. D. and Sun, C. K., "A Note on Exponential Absolute Stability", To be presented at the Fifth Asilomar Conference on Circuits and Systems, November, 1971.

[15] Eisenhart, L. P., Coordinate Geometry, Dover, New York, 1960.

[16] Dunn, W. R., Jr. and Chan, S. P., "Flowgraph Analysis of Linear Systems Using Remote Time-Shared Computation", Journal of the Franklin Institute, Vol. 288, No. 5, November, 1969.

```
      SUBROUTINE RESTNT(IN,VAL)
      IMPLICIT REAL*8(A-H,O-Z)
      COMMON/SHARE/X(100),DEL(100),A(100,100),N,M,MN,NP1,NM1
C THIS SUBROUTINE IS CALLED BY THE MAINLINE OF SUMT.'VAL' IS THE NUMERICAL VALUE
C OF CONSTRAINT NUMBERED 'IN'.EXCEPT FOR 'IN'=0 WHICH DENOTES THE FUNCTION TO BE
C MINIMISED.'N' DENOTES DIMENSION AND 'M' NUMBER OF CONSTRAINTS OF PROBLEM
      IF(IN)5,5,6
    5 VAL=-X(N)
      RETURN
    6 CALL LOOPER(IN,N,X,VAL)
      RETURN
      END
      SUBROUTINE LOOPER(IN,MX,R,VAL)
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION R(7),RA(6)
C THIS SUBROUTINE DETERMINES THE MINIMUM 'VAL' OF THE CONSTRAINT NUMBERED 'IN'.
C THE MINIMUM IS TAKEN OVER THE HYPERCUBE VERTICES AND EDGE/FACE MIDPOINTS.
C THE HYPERCUBE CENTER IS R(1),...,R(MX-1) AND EDGE IS R(MX).USING SIX NESTED DO
C LOOPS AND VARIABLE NESTING LOGIC,REGULATOR VECTOR UPTO SIX DIMENSION HANDLED.
      RO=R(MX)
      JM=MX-1
      VAL=1000.
      DO 51 IA=1,3
      RA(1)=R(1)-(IA-2)*RO
      IF(JM-1)17,17,18
   18 DO 52 JA=1,3
      RA(2)=R(2)-(JA-2)*RO
      IF(JM-2)17,17,19
   19 DO 53 KA=1,3
      RA(3)=R(3)-(KA-2)*RO
      IF(JM-3)17,17,20
   20 DO 54 LA=1,3
      RA(4)=R(4)-(LA-2)*RO
      IF(JM-4)17,17,21
   21 DO 55 MA=1,3
      RA(5)=R(5)-(MA-2)*RO
      IF(JM-5)17,17,22
   22 DO 56 NA=1,3
      RA(6)=R(6)-(NA-2)*RO
   17 CONTINUE
C SUBROUTINE CONSTR EVALUATES THE VALUE 'VAL' OF CONSTRAINT NUMBERED 'IN',GIVEN
C REGULATOR VECTOR 'RA' OF (MX-1) DIMENSION.
      CALL CONSTR(IN,RA,EVA)
      IF(EVA-VAL)5,5,6
    5 VAL=EVA
    6 CONTINUE
      GO TO (51,52,53,54,55,56),JM
   56 CONTINUE
   55 CONTINUE
   54 CONTINUE
   53 CONTINUE
   52 CONTINUE
   51 CONTINUE
      RETURN
      END
```

Fig. 1

SUBOPTIMAL COMPENSATOR DESIGN FOR THE LINEAR REGULATOR VIA SECOND-ORDER SENSITIVITY

W. A. BROWN
Monash University, Australia.
(on leave at the University of Waterloo, Canada)

ABSTRACT

The design of a compensation network for a linear system is treated as an optimal regulator problem in which a dynamic estimator for unavailable states is incorporated. Second-order sensitivity coefficients are derived relating changes in the quadratic performance index to variations in compensator parameters. A knowledge of these coefficients permits a systematic approach to the design of a suboptimal compensator of reduced order.

I. INTRODUCTION

The optimal solution to the linear qudratic regulator problem involves the instantaneous feedback of a linear combination of all system states. In practice however it is often undesirable or even impossible to measure all states directly. It then becomes necessary to incorporate a state estimator if full state variable feedback is to be retained. Luenberger has presented rules for the order of estimator required and Pearson[1] has presented a design technique for the overall optimization of the estimator and controller (the combination being known as a compensator). Pearson poses the overall optimization problem as a linear quadratic problem of order equal to the original system order plus the estimator (Luenberger) order. The feedback coefficient matrix for this higher order problem is used to calculate the coefficients of the compensator, and thereby to do away with feedback from the unavailable states.

For reasons of simplicity and economy it is often desirable to use a compensator of less than Luenberger order. Indeed experience indicates that systems of medium or high order can often be well controlled using relatively simple compensators (lead networks, proportional plus derivative controllers etc.). Thus what is needed is a systematic procedure for designing a compensator of order as low as possible, consistent with an acceptable level of performance deterioration. In this context the term 'suboptimal compensator' will be used to describe a compensator of less than Luenberger order.

The method outlined here for suboptimal compensator design employs a sensitivity analysis, to second-order accuracy, of the quadratic performance index to changes in compensator coefficients. It is a logical development of earlier work[2] on sensitivity analysis applied to suboptimal design of the linear regulator with incomplete state feedback. In that work it was shown that the calculation of second-order sensitivity coefficients about the optimum is computationally simple and provides a powerful basis for suboptimal design. Since Pearson's technique of compensator design involves the calculation of optimal feedback coefficients, the results for incomplete state feedback can be readily applied to the design of a reduced order compensator. Furthermore, because the sensitivity analysis is based upon the optimal solution, an immediate check on the magnitude of performance deterioration can be made at any stage in the suboptimal compensator design procedure.

In section II the development of equations for second-order sensitivity analysis is summarized and the problem of choice of initial conditions is discussed. In section III a procedure for reduced-order compensator design is developed, and in section IV an example is presented.

II. SECOND-ORDER SENSITIVITY ANALYSIS APPLIED TO CONTROLLER DESIGN

For the linear constant system

$$\dot{x} = Ax + Bu \tag{1}$$

with quadratic performance index

$$J = \int_0^\infty (x^T Qx + u^T Ru)\, dt \tag{2}$$

J is minimized when

$$u = -K^* x \tag{3}$$

where $$K^* = R^{-1} B^T P^* \tag{4}$$

$$J^* = x_o^T P^* x_o \tag{5}$$

and $$A^T P^* + P^* A - P^* BR^{-1} B^T P^* + Q = 0 \tag{6}$$

However for any feedback coefficient matrix, K, which is not necessarily optimal, the equations corresponding with (3), (5) and (6) are given by:

$$u = -Kx \tag{7}$$

$$J = x_o^T P x_o \tag{8}$$

and $$(A-BK)^T P + P(A-BK) + K^T RK + Q = 0 \tag{9}$$

providing the closed-loop system is asymptotically stable.

By successive differentiation of equation (9), first-, second- and higher-order sensitivity coefficient matrices relating incremental variations in P to incremental variations in K can be found. The first-order coefficient matrix for variations in K about K* must be zero as a consequence of the definition of optimality. It has been shown [2] that to second-order accuracy variations in P due to variations in K about K* ($\Delta K = K - K^*$) are given by:

$$cs(\Delta P) \approx -[I \times (A-BK^*)^T + (A-BK^*)^T \times I]\,[\Delta K^T \times \Delta K^T]\,cs(R) \qquad (10)$$

$$\text{Hence } \Delta J = x_o^T \, \Delta P \, x_o \qquad (11)$$

Equations (10) and (11) provide a detailed and reasonably accurate knowledge of performance sensitivity to control coefficient variations and therefore an excellent basis for suboptimal design.

At this stage it is important to note that the choice of initial conditions affects suboptimization. Whereas K* is unique in the sense that it does not depend upon x_o, if K is limited to a region not including K*, for every value of x_o there is a different value for the best K. Thus a procedure for suboptimal design must involve an assumption about the distribution of initial conditions. In practical applications a limited set of initial conditions is usually of prime importantance, or even a single set. Alternatively a distribution of initial conditions may be considered and an ensemble average employed [3].

In those situations where step response behaviour is considered important, one particular set of initial conditions is appropriate. Consider the original system, equation (1), with output equation $y = Cx$ and feedback equation $u = r - Kx$, which now includes a constant external input r. The response of the system for zero initial conditions is given by:

$$y = C\exp(A-BK)t\,(A-BK)^{-1}Br - C(A-BK)^{-1}Br \qquad (12)$$

The first term in this equation is the transient response and the second the steady state. If r is chosen so that y has a desired steady state level (typically unity), then the transient part is the error and is identical with a zero-input response with initial conditions given by

$$x_o = (A-BK)^{-1}Br \qquad (13)$$

Thus a suboptimal design based upon this set of initial conditions is equivalent to design for a step response. (In the multivariable case a number of initial condition sets may be employed, each one corresponding with step response at a particular output.)

Once an assumption about initial conditions has been made, equations (10) and (11) can be manipulated to form:

$$\Delta J \approx (rs\,\Delta K)\,S\,(rs\,\Delta K)^T \qquad (14)$$

where S is a matrix of second-order sensitivity coefficients, dependent upon x_o or a distribution of x_o. Equation (14) provides a means (a) to find which elements of K have little influence on J and can therefore be safely reduced to zero and (b) having fixed certain elements of ΔK accordingly, to re-adjust the remaining elements of ΔK to minimize ΔJ. Step (b) involves solution of a simple quadratic minimization problem and ensures that the suboptimal system is as good as it can be.

III. A PROCEDURE FOR SUBOPTIMAL COMPENSATOR DESIGN

Because the Pearson compensator design procedure employs optimal coefficients it is amenable to the sensitivity analysis outline above. The following is an outline of Pearson's procedure especially adapted to suit the design of a reduced-order compensator. It is developed for a single-input single-output system, but can be extended to the multivariable case [4].

Consider the plant of equation (1) where x has n elements, u is a scalar and the output is also a scalar given by

$$y = Cx \tag{15}$$

Whilst only one output is available from the plant, it is assumed temporarily that p internal states are also available, where $p \leq n-2$. (The feedback coefficients from these states will later be reduced to zero.) These states must be carefully chosen and without loss of generality can be taken as the last p states of x.

Effectively p+1 states are available from the plant, and therefore n-p-1 further states must be estimated. (It is assumed the first n-p states are observable from y.) Thus a compensator of order n-p-1 is required. Following Pearson, the output of a chain of n-p-1 integrators is connected to u. Thus the compensator state equations are given by:

$$\dot{v} = Dv + Ew \tag{16}$$

$$u = Hv \tag{17}$$

where v is the compensator state, w is the scalar input to the compensator and

$$D = \begin{bmatrix} 0 & 1 & \cdots & & 0 \\ 0 & 0 & \cdot & & \cdot \\ \cdot & & \cdot & & 1 \\ \cdot & & & \cdot & \\ 0 & \cdots & & \cdot & 0 \end{bmatrix} \qquad E = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{bmatrix}$$

$$H = [1 \quad 0 \quad \cdots \quad 0]$$

An overall optimization problem for the combined system (order 2n-p-1) is now solved to obtain optimal feedback coefficients from all states of the plant and compensator. The $\bar{A}$, $\bar{B}$, $\bar{Q}$ and $\bar{R}$ matrices for this problem are given by:

$$\bar{A} = \begin{bmatrix} A & BH \\ 0 & D \end{bmatrix} \qquad \bar{B} = \begin{bmatrix} 0 \\ E \end{bmatrix}$$

$$\bar{Q} = \begin{bmatrix} Q & 0 \\ 0 & (H^T RH + Q') \end{bmatrix}$$

and $\bar{R}$ is carefully chosen to weight the effect of w. Q' must also be carefully chosen for a satisfactory compensator design.

The solution to the optimization problem is given by:

$$w = -[F^* \quad G^* \quad L^*] \begin{bmatrix} x \\ v \end{bmatrix} \tag{18}$$

where F* has n-p elements, G* has p and L* has n-p-1. That is, the elements of G* are the feedback coefficients from the last p states of x.

In order to eliminate the need for the first n-p states of x, an equivalent form for equation (18) can be found as follows:

$$w = -[\Phi^* \quad \Gamma^* \quad \Lambda^*] \begin{bmatrix} y \\ y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(n-p-1)} \\ x_{n-p+1} \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ v \end{bmatrix} \tag{19}$$

where the feedback coefficient matrix is partitioned in the same way as equation (18). Obviously $\Gamma^* = G^*$ and $\Lambda^* = L^*$. However Φ^* is linearly related to all three matrices, F*, G* and L*. It is convenient to summarize the relationship in the linear transformation:

$$[F^* \quad G^* \quad L^*] = [\Phi^* \quad \Gamma^* \quad \Lambda^*]T \tag{20}$$

where

$$T = \begin{array}{c} \\ \\ n-p \\ \\ \\ n-1 \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} n & \qquad\qquad\qquad n-p-1 \end{array} \\ \left[\begin{array}{c|ccccc} C & 0 & 0 & \cdots & \cdots & 0 \\ CA & CB & 0 & & & \vdots \\ \vdots & \vdots & & & & \vdots \\ CA^{n-p-1} & CA^{n-p-2}B & \cdots & \cdots & \cdots & CB \\ \hdashline 0 & & & I & & \end{array}\right] \\ \begin{array}{cc} n-p & \qquad\qquad\qquad n-1 \end{array} \end{array} \tag{21}$$

The optimal compensator transfer function is given by:

$$u = -H[sI - D + E\Lambda^*]^{-1} E\{ \sum_{i=1}^{n-p} \Phi_i^* s^{i-1} y + \sum_{j=1}^{p} \Gamma_j^* x_{n-p+j} \} \tag{22}$$

The sensitivity analysis developed in section II can now be applied to analyse the effect of reducing Γ to zero and to suboptimize Λ and Φ, the compensator coefficients. Equation (20) will be valid away from the optimum, thus:

$$[F \quad G \quad L] = [\Phi \quad \Gamma \quad \Lambda]T \tag{23}$$

$$[\Delta F \quad \Delta G \quad \Delta L] = [\Delta\Phi \quad \Delta\Gamma \quad \Delta\Lambda]T \tag{24}$$

and equation (14) becomes

$$\Delta J \approx [\Delta\Phi \quad \Delta\Gamma \quad \Delta\Lambda]T\,S\,T^T [\Delta\Phi \quad \Delta\Gamma \quad \Delta\Lambda]^T \tag{25}$$

The design procedure is then to choose the elements of $\Delta\Gamma$ so that Γ becomes zero, and minimize ΔJ with respect to $\Delta\Phi$ and $\Delta\Lambda$. In this way feedback from the last p states of x is eliminated and the compensator is suboptimized. The design cycle is completed by checking the final ΔJ against J^* to see that performance deterioration caused by use of a reduced order compensator is within an acceptable limit.

IV. EXAMPLE

Consider the following system design problem (See [2]):

$$A = \begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1.5 \end{matrix} \qquad B = \begin{matrix} 0 \\ 0 \\ 1 \end{matrix} \qquad Q = \begin{matrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}$$

$$C = \begin{matrix} 1 & 0 & 0 \end{matrix} \qquad\qquad R = 1$$

For this third-order system, a second-order compensator would normally be needed. But a first-order compensator will be considered and it will be assumed initially that x_3 is available (p=1).

Taking $Q' = 0$ and $\bar{R} = 1$, the optimization problem becomes:

$$\bar{A} = \begin{matrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1.5 & 1 \\ 0 & 0 & 0 & 0 \end{matrix} \qquad \bar{B} = \begin{matrix} 0 \\ 0 \\ 0 \\ 1 \end{matrix} \qquad \bar{Q} = \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{matrix}$$

For this problem T = I and the solution is given by:

$$F^* = [1.00 \quad 3.06] \quad = \quad \Phi^*$$

$$G^* = 1.61 \quad = \quad \Gamma^*$$

$$L^* = 2.06 \quad = \quad \Lambda^*$$

The system will be suboptimized for a step resonse, for which the equivalent initial conditions (see equation (13)) are given by $x_o = [1 \quad 0 \quad 0]$. A second-order sensitivity analysis for this problem yields:

$$\Delta J \approx [\Delta\Phi \quad \Delta\Gamma \quad \Delta\Lambda] \begin{bmatrix} 2.57 & -0.50 & -0.23 & -0.35 \\ -0.50 & 0.23 & 0 & -0.06 \\ -0.23 & 0 & 0.03 & 0.09 \\ -0.35 & -0.06 & 0.09 & 0.19 \end{bmatrix} [\Delta\Phi \quad \Delta\Gamma \quad \Delta\Lambda]^T$$

Thus setting $\Delta\Gamma = -1.6$ and minimizing ΔJ with respect to the remaining compensator coefficients yields

$$\Delta\Phi = [0.16 \quad 0.69] \qquad \Delta\Lambda = 1.22$$

$$\Phi = [1.16 \quad 3.75] \qquad \Gamma = 0 \qquad \Lambda = 3.28$$

The final system is shown in Fig. 1 and is seen to have a simple lead network compensator in the feedback path. Its step resonse is shown in Fig. 2 together with the step response of the system with full state variable feedback.

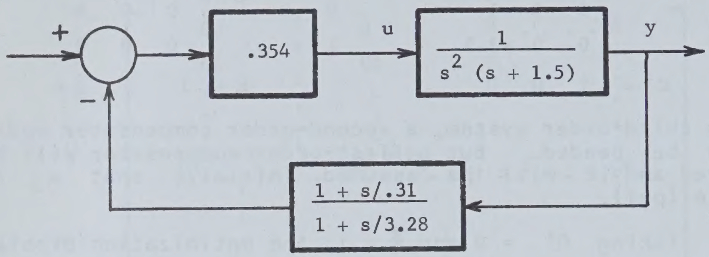


Figure 1. Feedback system with suboptimal compensator.

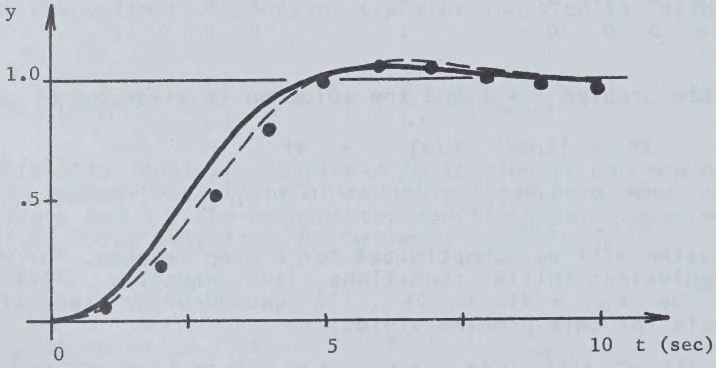


Figure 2. Step response for the system with optimal state-variable feedback (full line), optimal first order compensator and x_3 included (broken line), and the suboptimal compensator (dotted line).

V. CONCLUSION

It has been shown how a compensator of low order can be designed for a linear system using an analytical approach. The only 'guesswork' involved is the choice of performance index weighting coefficients and the arrangement of A and B plant matrices best suited to the suboptimization procedure. The major attractions of the method are (a) the suboptimization is made with respect to the optimal solution so that a direct comparison can be made during the design procedure and (b) that the S matrix (or TST^T) provides a complete picture of control coefficient sensitivity.

ACKNOWLEDGEMENT

The author is indebted to the University of Waterloo for making this work possible especially Professor A. R. M. Noton, and to Professor W. J. Vetter for helpful discussions.

REFERENCES

[1] Pearson, J. B., "Compensator design for dynamic optimization", International Journal of Control, 1969, Vol. 9, No. 4, p. 473.

[2] Brown, W. A. and Vetter, W. J., "Suboptimal design of the linear regulator with incomplete state feedback via second order sensitivity", International Journal of Control, to be published.

[3] Dabke, K. P., "Linear control with incomplete state feedback and known initial-state statistics", International Journal of Control, 1970, Vol. 11, No. 1, p. 133.

[4] Pearson, J. B. and Ding, C. Y., "Compensator Design for multivariable linear systems", IEEE Transactions on Automatic Control, 1969, Vol. AC-14, No. 2, p. 130.

AN IMPROVED DESIGN TECHNIQUE FOR PARAMETER ADAPTIVE CONTROL SYSTEMS

J. W. GILBART
Resident Research Associate
Goddard Space Flight Center
Greenbelt, Maryland

R. V. MONOPOLI AND E. LIZEWSKI
University of Massachusetts
Amherst, Massachusetts

ABSTRACT

A parameter adaptive control technique is developed utilizing Liapunov's Direct Method. The design approach represents an improvement over previous methods in that only one derivative network is required for controller implementation. Furthermore, the technique may be applied to plants with right half plane zeros while maintaining a bounded control input.

I. Introduction

The imposition of stringent performance requirements on systems which undergo wide parameter variations frequently necessitates the use of an adaptive control system. For example, adaptive techniques have been applied to devices such as autopilots [1] where satisfactory response characteristics are required in the presence of unpredictable changes in flight conditions along an aircraft's trajectory. More recently, [2], a parameter adaptive system has been proposed for the control of an optical tracking mount in order to maintain precise pointing accuracy while certain mount parameters undergo wide variations.

Early designs for these systems were based on gradient techniques used to minimize a suitable performance index. The most prominent of these was a model referenced parameter adaptive scheme, called the "M.I.T. rule" which was developed in [3]. It was shown in [4], however, that analysis difficulties as well as possible instability resulted from this approach. As an alternative method a design technique was suggested in [4] and subsequently modified and extended in [5] - [11], that utilizes Liapunov's direct method as the basis for synthesizing the adaptive loops. This approach, however, has a severe practical limitation in that, in general, it requires n plant states or n-1 derivatives of the plant output (n^{th} order plant) if the plant states are in phase variable form. If the plant contained m zeros, it was shown in [9] that this requirement could be reduced to n-m-1 derivative networks. Another disadvantage of the Liapunov approach, as shown in [9] and [12], was the fact that an unbounded control input resulted when it was applied to non-minimum phase plants; i.e., plants whose zeros lie in the right half s-plane. The transfer function relating pitch motion to command input for an airframe, for example, frequently contains a right half plane zero [12].

In this paper, a design scheme is presented which alleviates these practical disadvantages. Adaptive laws are derived which necessitate only one derivative network for their implementation, regardless of the number of plant zeros and poles. Furthermore, contrary to previous designs, the technique in this paper can be applied to non-minimum phase plants while maintaining a bounded plant input.

II. Preliminary Definitions

The class of plants to which the technique applies can be described by

$$\sum_{i=0}^{n} a_{ip}\, x_p^{(n-i)} = u^{(m)} + \sum_{i=1}^{m} b_{ip} u^{(m-i)} \qquad (1)$$

where the a_{ip}, i=0, n and b_{ip}, i=1, m ($m \leq n-1$) are unknown parameters which are constant with respect to the adaptation period of the adjustable parameters. The plant output is denoted by $x_p(t)$, $x_p^{(i)}(t)$ represents the i^{th} derivative of $x_p(t)$ with $x_p^{(0)}(t) = x_p(t)$, and u(t) is the control input through which the adjustable parameters are to be introduced. The plant is assumed stable but any or all of the roots of

$$s^m + \sum_{i=1}^{m} b_{ip} s^{(m-i)} = 0$$

may lie in the right half s-plane. It is assumed that the desired plant characteristics are represented by a stable model reference system described by

$$\sum_{i=0}^{n} a_{id}\, x_d^{(n-i)} = r^{(m)} + \sum_{i=1}^{m} b_{id} r^{(m-i)} \qquad (2)$$

where $x_d(t)$ is the desired plant response, r(t) is the reference input, and the parameters a_{id}, i=0, n and b_{id}, i=1, m are known constants.

The philosophy of this adaptive scheme, while employing ideas similar to those in [4] - [11], is basically different. Here, as shown in Fig. 1, the design is based on an input error signal, $e_u(t) = u_0(t) - u(t)$, rather than the output error signal, $e(t) = x_d(t) - x_p(t)$ employed in previous designs. In the case where the plant is minimum phase, the signal $u_o(t)$ is to be generated in a manner such that the steady state condition $e_u(t) = 0$ implies that e(t) = 0, i.e., $x_p(t) = x_d(t)$. Liapunov theory is then employed to guarantee that $\lim_{t \to \infty} e_u(t) = 0$. For non-minimum phase plants, $u_o(t)$ is formulated so that $e_u(t) = 0$ implies that the characteristic response of the compensated plant is the same as the characteristic response of the model. In this instance, adjustable parameters are introduced to compensate for the plant poles only.

Before the design of the adaptive filter and adjustable compensator is considered, some additional definitions are in order. As shown in Fig. 2, the plant may be represented by its linear, time-invariant transfer function, (neglecting initial conditions for simplicity), given by $X_p(s)/U(s) = M_p(s)/N_p(s)$, where

$$M_p(s) = s^m + \sum_{i=1}^{m} b_{ip} s^{(m-i)}$$

$$N_p(s) = \sum_{i=0}^{n} a_{ip} s^{(n-i)} \tag{3}$$

The unknown constants a_{ip}, i=0, n and b_{ip}, i=1, m are defined in eqn. (1). Likewise, the transfer function of the reference model is given by $X_d(s)/R(s) = M_d(s)/N_d(s)$ where $M_d(s)$ and $N_d(s)$ are defined in a manner similar to eqn. (3). Furthermore, it is seen in Fig. 2 that the signals u(t) and $u_o(t)$ are generated by passing the as yet undefined signals w(t) and y(t), respectively, through the linear, time invariant network with transfer function $F(s)/(N_d(s) + K)$. F(s) is any nth order polynomial defined by

$$F(s) = s^n + \sum_{i=1}^{n} f_i s^{(n-i)} \tag{4}$$

where the f_i, i=1, n are positive constants and the roots of F(s) = 0 lie in the open left half s-plane with at least one root being real and negative. The positive constant, K, is chosen small enough so that the roots of $N_d(s) + K = 0$ are in the left half s-plane.

III. Development

The formulation of a first order differential equation in the input error, $e_u(t)$, is now considered. From Fig. 2, it is easily seen that

$$E_u(s)[N_d(s) + K] = F(s)[Y(s) - W(s)] + KU(s) \tag{5}$$

and that

$$U(s) = N_p(s)X_p(s) - [M_p(s) - 1]U(s) \tag{6}$$

If eqn. (6) is substituted into eqn. (5), the result is

$$E_u(s)[N_d(s) + K] = F^1(s)(s+c)[Y(s) - W(s)] + K[N_p(s)X_p(s) - [M_p(s) - 1]U(s)] \tag{7}$$

where the polynomial F(s) has been factored as follows:

$$F(s) = F^1(s)\,(s+c) \tag{8}$$

In eqn. (8), $F^1(s)$ is of order n-1 and c is real and positive. If both sides of eqn. (7) are divided by $F^1(s)$ and the Laplace inverse is taken, it is shown in [9] that eqn. (9) results if an appropriate choice of F(s) is made.

$$K(a_{od}\dot{e}_u + d_1 e_u) = \dot{y} - \dot{w} + c(y - w) + d_2 e_u^* - Ku_f + \sum_{i=1}^{n+m+1} \gamma_i z_i \qquad (9)$$

This procedure is outlined in detail in [9] and will not be discussed here. In eqn. (9), u_f is defined by $u_f = \mathcal{L}^{-1}\left\{\frac{s^m U(s)}{F'(s)}\right\}$, d_1 and d_2 are known positive constants, the γ_i, i=1, n+m+1 are unknown constants, and the z_i, i=1, n+m+1 are signals generated from $\dot{x}_p$, x_p, u, and e_u and filtered versions of these signals. For example, in the case of a second order system with one zero; i.e., n=2, m=1, and $F(s) = (s+c)^2$, the parameters d_1, d_2, γ_i, i=1, 4, and the signals z_i, i=1, 4 are defined by $d_1 = a_{id} - a_{od}c$, $d_2 = c(Ka_{id} - cKa_{od}) - Ka_{2d}$, $\gamma_1 = Ka_{op}$, $\gamma_2 = Ka_{1p} - Ka_{op}c$, $\gamma_3 = Ka_{2p} - c(Ka_{1p} - Ka_{op}c)$, $\gamma_4 = -K(b_{1p} - 1)$, $z_1 = \dot{x}_p$, $z_2 = x_p$, $z_3 = x_p^*$, $z_4 = u^*$. The * operation is defined by

$$g^*(t) = \mathcal{L}^{-1}\left\{\frac{G(s)}{s+c}\right\} \qquad (10)$$

The signal w(t) is now selected as

$$w = y + d_2(e_u)^* - Ku_f^* + v^* \qquad (11)$$

where

$$v(t) = \sum_{i=1}^{n+m+1} k_i z_i \qquad (12)$$

In eqn. (12), the k_i, i=1, n+m+1 are adjustable parameters introduced to compensate for the unknown parameters γ_i, i=1, n+m+1, respectively. When eqn. (12) is substituted into eqn. (9), the result is

$$K(a_{od}\dot{e}_u + d_1 e_u) = \sum_{i=1}^{n+m+1} \Delta_i z_i \qquad (13)$$

where $\Delta_i = \gamma_i - k_i$, i=1, n+m+1.

The next step in the design procedure is to define adaptive laws; i.e., definitions for the k_i, i=1, n+m+1, in order to assure that $\lim_{t\to\infty} e_u(t) = 0$. For this purpose the following Liapunov function is selected:

$$V = \frac{1}{2}\left(e_u^2 + \sum_{i=1}^{n+m+1} \lambda_i \Delta_i^2\right) \qquad (14)$$

In eqn. (14), the λ_i, i=1, n+m+1 are arbitrary positive constants. If the adaptive laws are selected as

$$k_i(t) = \frac{1}{\lambda_i}\int_0^t e_u z_i\,dt + \frac{\alpha}{\lambda_i}\int_0^t (\dot{e}_u + d_1 e_u) z_i\,dt \qquad (15)$$

where α is an arbitrary positive constant, then the time derivative of (14) becomes

$$\dot{V} = -\left(\frac{d_1}{Ka_{od}}\right)e_u^2 - \left(\frac{\alpha}{Ka_{od}}\right)\left(\sum_{i=1}^{n+m+1} \Delta_i z_i\right)^2 \tag{16}$$

The second term in (15) is new here and is introduced in order that $\dot{V}$ in eqn. (16) will contain a negative semidefinite form in the parameter differences, Δ_i, i=1, n+m+1. Previous formulations did not penalize these parameter differences in this manner and, hence, either they did not converge to zero at all, as in the case of incomplete adaptation due to frequency poor inputs, or their convergence was slow, as noted in [9]. In this regard, a theorem developed in [13] can be utilized which guarantees that $\lim_{t\to\infty} \Delta_i(t)=0$ for all i if the input r(t) contains at least $\frac{n+m+1}{2}$ different frequencies, none of which undergo a phase shift of exactly $k\pi$, $k=\pm 1, \pm 2, \ldots$, at the plant output.

Finally, it remains to choose the signal, y(t), in an appropriate manner so that $\lim_{t\to\infty} \Delta_i(t)=0$, i=1, n+m+1 implies that $\lim_{t\to\infty} e(t)=0$; i.e., $x_p(t) = x_d(t)$. Utilizing the steady state condition $E_u(s)=0$ in eqn. (7) and solving for W(s) yields

$$W(s) = Y(s) + \frac{KN_p(s)X_p(s)}{F(s)} - \frac{K[M_p(s) - 1]U(s)}{F(s)} \tag{17}$$

Since, from Fig. 2, the signals $x_p(t)$ and u(t) can be defined by their Laplace transforms as

$$X_p(s) = M_p(s)U(s)/N_p(s) \tag{18}$$

$$U(s) = F(s)W(s)/[N_d(s) + K] \tag{19}$$

eqns. (17), (18), and (19) may be solved to yield

$$X_p(s)/Y(s) = F(s)M_p(s)/N_p(s)N_d(s) \tag{20}$$

Hence, in order that the steady state condition $E_u(s)=0$ implies that $X_p(s) = X_d(s)$, Y(s)/R(s) must be given by

$$Y(s)/R(s) = N_p(s)M_d(s)/F(s)M_p(s) \tag{21}$$

However, if $M_p(s)$ contains right half plane zeros, then y(t), as defined by eqn. (21), will be unbounded. In this case it is desirable that Y(s)/R(s) be defined by

$$Y(s)/R(s) = N_p(s)/F(s) \tag{22}$$

From eqns. (20) and (22), $X_p(s)$ then becomes

$$X_p(s) = M_p(s)R(s)/N_d(s) \tag{23}$$

i.e., compensation is provided for plant poles only.

In order to define the signal $y(t)$ so that its steady-state transfer function is given by eqn. (21) (if the plant has no right half plane zeros) or by eqn. (22) (if the plant has right half plane zeros), it is necessary to examine the steady-state values of the $k_i(t)$, $i=1, n+m+1$, and then define $y(t)$ to be a function of these adjustable parameters accordingly. For example, in the case of a second order plant with no zeros, i.e., $M_d(s) = s + b_{1d}$, $N_p(s) = a_{op}s^2 + a_{1p}s + a_{2p}$, $F(s) = (s + c)^2$, and $M_p(s) = s + b_{1p}$, eqn. (21) becomes

$$Y(s)/R(s) = \frac{(s + b_{1d})(a_{op}s^2 + a_{1p}s + a_{2p})}{(s + b_{1p})(s + c)^2} \qquad (24)$$

From the definitions of the γ_i, $i=1, n+m+1$ in eqn. (9) and the fact that

$$\lim_{t\to\infty} k_i(t) = \gamma_i, \; i=1, n+m+1 \qquad (25)$$

it is easily seen that if $y(t)$ is defined by eqn. (26), then, after adaptation is complete, $y(t)$ will be related to $r(t)$ by eqn. (24).

$$y(t) = (b_{1d} - 1 + \frac{k_4}{K})y_f + \frac{1}{K}[k_1\dot{r}^* + k_2r^* + k_3(r^*)^*] \qquad (26)$$

If, on the other hand, $b_{1p} < 0$, $y(t)$ defined by eqn. (27),

$$y(t) = \frac{1}{K}[k_1\dot{r}^* + k_2r^* + k_3(r^*)^*] \qquad (27)$$

yields the desired result (eqn. (22)) after adaptation is complete.

IV. Conclusions

A parameter adaptive control technique, utilizing only a first order derivative network for controller implementation, has been presented. In addition, the technique applies to plants whose zeros lie in the right half plane. In this instance, compensation is provided for the plant poles only in order to maintain a bounded control input.

ACKNOWLEDGEMENT

Partial support of this research was obtained from the National Aeronautics and Space Administration under Grant No. NGL-22-010-018. Dr. James W. Gilbart was in the Dept. of Elec. Engr. at the University of Massachusetts when most of this work was done.

REFERENCES

1. Mellen, D.L., "Applications of Adaptive Flight Control", Proceedings of the First International Symposium on Optimizing and Adaptive Control, Rome, Italy, 1962, pp. 163-175.

2. Gilbart, J.W., "The Model Referenced Adaptive Control of an Optical Tracking Mount," NASA X-Document #525-71-205, June, 1971.

3. Osburn, P.V., Whitaker, H.P., and Kezar, A., "New Developments in the Design of Adaptive Control Systems", Inst. of Aeronautical Sciences, paper 61-39.

4. Parks, P.C., "Liapunov Redesign of Model Reference Adaptive Control Systems", I.E.E.E. Transactions on Automatic Control, July, 1966, pp. 362-367.

5. Monopoli, R.V., "Liapunov's Method for Adaptive Control System Design", I.E.E.E. Transactions on Automatic Control, October, 1967, Vol. AC-12, No. 5, p. 625.

6. Monopoli, R.V., Gilbart, J.W., and Thayer, W.D., "Model Reference Adaptive Control Based on Liapunov-Like Techniques", Proceedings of the Second IFAC Symposium on System Sensitivity and Adaptivity, pp. F.24 - F.31, August, 1968.

7. Winsor, C.A. and Roy, R.J., "Design of Model Reference Adaptive Control Systems by Liapunov's Second Method," I.E.E.E. Transactions on Automatic Control, April, 1968, Vol. AC-13, No. 2, p. 204.

8. Phillipson, P.H., "Design Methods for Model-Reference Adaptive Systems," Proceedings of the Institution of Mechanical Engineers, 1968-1969, Vol. 183, Part 1.

9. Monopoli, R.V., and Gilbart, J.W., "Model Reference Adaptive Control Systems with Feedback and Prefilter Adjustable Gains", Proceedings of the Fourth Annual Princeton Conference on Information Sciences and Systems, March, 1970.

10. Gilbart, J.W., Monopoli, R.V., and Price, C.F., "Improved Convergence and Increased Flexibility in the Design of Model Reference Adaptive Control Systems", Proceedings of the 1970 I.E.E.E. Symposium on Adaptive Processes (9th): Decision and Control, December, 1970.

11. Gilbart, J.W. and Monopoli, R.V., "Parameter Adaptive Control of a Class of Time-Varying Plants," Proceedings of the Fifth Annual Princeton Conference on Information Sciences and Systems, March 1971.

12. Price, C.F. and Koenigsberg, W.D., "Adaptive Control and Guidance for Tactical Missiles," Contract Report for the Department of Navy, Contract #N00014-69-C-0391, The Analytic Sciences Corporation, Reading, Massachusetts.

13. Lyon, P.M., "Rapid Identification of Linear and Nonlinear Systems," AIAA Journal, Vol. 5, No. 10, October, 1967.

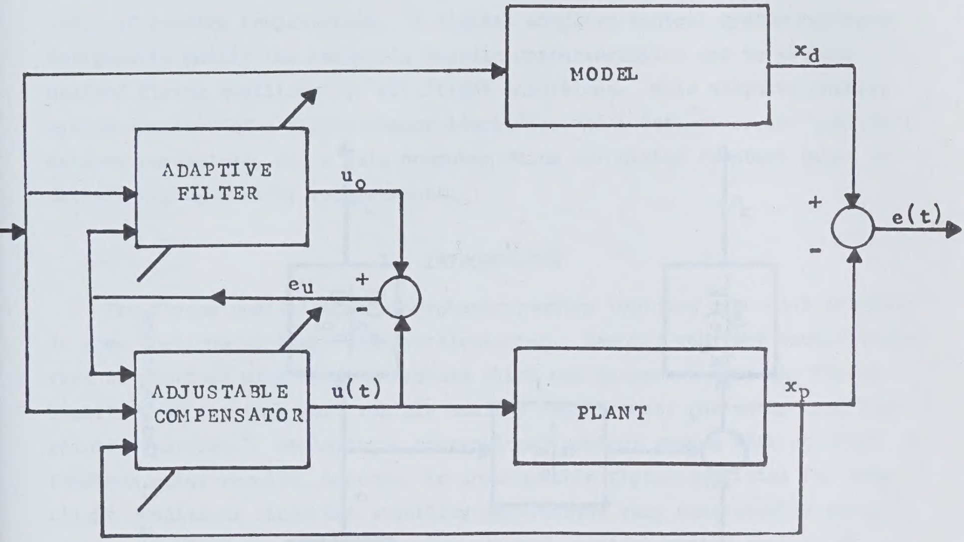
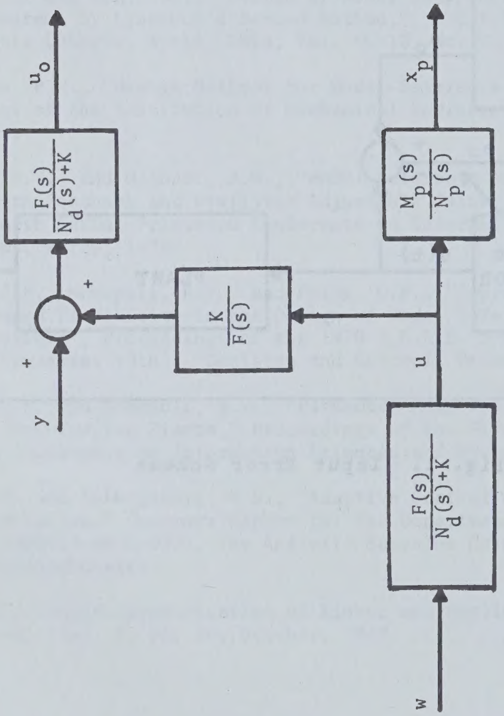

MODEL
x_d
ADAPTIVE FILTER
u_o
+
$e(t)$
-
e_u
+
-
ADJUSTABLE COMPENSATOR
$u(t)$
PLANT
x_p

Fig. 1. Input Error Scheme



y
+
+
u_o
$\frac{F(s)}{N_d(s)+K}$
$\frac{K}{F(s)}$
u
w
$\frac{F(s)}{N_d(s)+K}$
$\frac{M_p(s)}{N_p(s)}$
x_p

Fig. 2: Design Scheme

DESIGN OF A DIGITAL ADAPTIVE CONTROL SYSTEM FOR REENTRY VEHICLES

Jose L. Picon-Jimenez, Research Associate, Langley Research Center,
Hampton, Virginia,
Dr. R. C. Montgomery, Aerospace Technologist, Langley Research Center,
Hampton, Virginia, and
Dr. L. L. Grigsby, Department of Electrical Engineering, VPISU,
Blacksburg, Virginia

ABSTRACT

The flying qualities of atmospheric reentry vehicles experience considerable variations due to the wide changes in flight conditions characteristic of reentry trajectories. A digital adaptive control system has been designed to modify the vehicle's dynamic characteristics and to provide desired flying qualities for all flight conditions. This adaptive control system consists of a finite-memory identifier which determines the vehicle's unknown parameters, and a gain computer which calculates feedback gains to satisfy flying quality requirements.

I. INTRODUCTION

The flying qualities of atmospheric reentry vehicles are unsatisfactory in some portions of their entry trajectories. Reentry vehicles should therefore be provided with control systems which can guarantee that the flying qualities of the augmented vehicle satisfy requirements currently used for piloted vehicles.[1] Employing a conventional control system with constant feedback gains results, however, in unacceptable flying qualities for some flight conditions since the stability derivatives vary considerably along their entry trajectories. Hence, an adaptive control system capable of yielding satisfactory flying qualities for all flight conditions is needed for this type of vehicle.

This paper presents a digital adaptive control system which has been designed to guarantee that the flying qualities of the vehicle satisfy ideal requirements throughout the entire range of flying conditions. This adaptive control system consists of a finite-memory identifier which determines the vehicle's stability derivatives, and a gain computer which checks the vehicle's flying qualities to determine the necessary feedback gains modifications.

The finite-memory identifier determines the stability derivatives using a linear model of the nonlinear equations of motion whose unknown parameters are the stability derivatives, and a shifting memory record which stores only the most recent vehicle's measurements. The model's parameters are adjusted on-line by means of a modified Newton-Raphson method which seeks a least-squares fit between the model's response and the stored measurements. The finite-memory identifier has been designed to avoid the divergence problems characteristic of Kalman filters when modeling errors are large compared with noise inputs.[2,3] Jazwinski has shown that, in such cases, finite-memory estimators are less subject to modeling errors than Kalman filters.[4]

Once the stability derivatives have been identified, the vehicle flying qualities are determined from simple algebraic relations.[1] The flying qualities are then compared with given ideal specifications to decide whether the feedback gains employed in the vehicle must be modified. New feedback gains are calculated if necessary using linear control theory.

II. VEHICLE'S DYNAMICS AND DESIGN REQUIREMENTS

The longitudinal dynamics of aircraft are discussed in Ref. 5. The differential equations which describe these dynamics are written below with the stability axes as a frame of reference (Fig. 1):

$$\frac{dV}{dt} = - \bar{q} \frac{S}{m} C_D - g \sin \gamma \tag{1.a}$$

$$\frac{d\gamma}{dt} = \bar{q} \frac{S}{mV} C_L - \frac{g \cos \gamma}{V} \tag{1.b}$$

$$\frac{dq}{dt} = \bar{q} \frac{S\bar{c}}{J_{yy}} C_M \tag{1.c}$$

$$\frac{d\theta}{dt} = q \tag{1.d}$$

$$\frac{dh}{dt} = V \sin \gamma \tag{1.e}$$

where V is the air velocity, $\bar{q}$ the dynamic pressure, S the wing area, m the vehicle's mass, C_D the drag coefficient, g earth gravity, γ the flight-path angle, C_L the lift coefficient, $\bar{c}$ the chord, J_{yy} the moment of inertia about the pitch axis, q the pitch rate, θ the pitch angle, and h the altitude. Also,

$$C_M = C_m + C_{m_q} \frac{\bar{c}}{2V} q, \tag{2}$$

where C_m is the static pitching-moment coefficient and C_{m_q} the pitch-damping coefficient. The aerodynamic coefficients C_D, C_L, C_m, and C_{m_q} are nonlinear functions of Mach number M, angle of attack α, and elevator deflection δ_e. They are shown as a function of angle of attack for various flight conditions in Figures 2 through 4, for a vehicle with the following characteristics: $S = 14.6\ m^2$, $\bar{c} = 64.6\ m$, $J_{yy} = 6404.88\ kg\ m^2$, $m = 2494.76\ kg$.

The vehicle's response to an elevator step of 5^o is shown in Figure 5. It illustrates two predominant modes: a long-period mode and a short-period mode. The flying quality requirements employed in this paper are based on the characteristics of the vehicle's short-period mode.

Design Requirements

The characteristics of the short-period mode are determined by linearizing (1.b) through (1.d) about some constant nominal conditions:

$$\begin{aligned}\frac{d}{dt}\Delta\gamma &= \frac{1}{V_{nom}}\left(-\bar{q}_{nom}\frac{S}{m}C_{L_\alpha} + g\sin\gamma_{nom}\right)\Delta\gamma + \bar{q}_{nom}\frac{S}{mV_{nom}}C_{L_\alpha}\Delta\theta \\ &\quad + \frac{1}{V_{nom}}\left(q_{nom}\frac{S}{m}C_{L_0} - g\cos\gamma_{nom}\right) \qquad (3)\\ \frac{d}{dt}\Delta q &= -\bar{q}_{nom}\frac{S\bar{c}}{J_{yy}}\left(-C_{m_\alpha}\Delta\gamma + C_{m_q}\frac{\bar{c}}{2V_{nom}}\Delta q + C_{m_\alpha}\Delta\theta + C_{m_{\delta_e}}\Delta\delta_e + C_{m_0}\right)\\ \frac{d}{dt}\Delta\theta &= \Delta q + q_{nom}\end{aligned}$$

where the off-trim coefficients C_{L_0} and C_{m_0}, as well as the stability derivatives C_{L_α}, C_{m_α}, $C_{m_{\delta_e}}$, and C_{m_q} are evaluated at the chosen nominal conditions.

The flying qualities employed in this paper are based on the characteristics of (3), namely, its damping ratio ζ_{SP} and undamped natural frequency versus normal acceleration per angle of attack $\frac{\omega_{SP}}{\sqrt{(n/\alpha)}}$. Previous aircraft experience indicates that appropriate values of these flying qualities are:[7]

$$\begin{aligned}&0.35 \le \zeta_{SP} \le 1.3\\ &0.30 \le \frac{\omega_{SP}}{\sqrt{(n/\alpha)}} \le 2.1\end{aligned} \qquad (6)$$

where

$$(n/\alpha) = C_{L_\alpha}\frac{S}{mg}\bar{q}. \qquad (7)$$

Figure 6 illustrates how the flying qualities of the uncontrolled vehicle vary along a typical reentry trajectory. It indicates that the vehicle's damping is too low for all flight conditions, and that $\omega_{SP} / (n/\alpha)$ fails to satisfy requirements (6) at large values of altitude and speed. The vehicle must therefore be provided with a control system to guarantee that requirements (6) are satisfied. Since the vehicle's stability derivatives do not remain constant throughout the vehicle's flight, an adaptive control system should be useful for this vehicle.

III. ADAPTIVE CONTROL SYSTEM DESCRIPTION

General Description

The adaptive control system presented here consists of a state reconstructor and an airborne computer to identify the vehicle's parameters and provide suitable feedback gains. During flight, measurements from axial and normal acceleration, angle of attack, dynamic pressure, flight-path angle, and pitch-rate sensors are employed in the state reconstructor to obtain airspeed and pitch angle. The vehicle's parameters are then identified employing a finite-memory identifier whose inputs are dynamic pressure, flight-path angle, stick deflection, and pitch-rate measurements, as well as filtered airspeed and pitch angle. Based on the identified stability derivatives, a gain computer determines feedback gain modifications. A schematic diagram of the overall system is exhibited in Figure 7.

State Reconstructor

Since the system parameters are unknown, an observer cannot be constructed to obtain estimates of pitch rate and airspeed. Hence a state reconstructor has been used to obtain airspeed and pitch angle from available measurements of axial and normal acceleration, dynamic pressure, angle of attack, and pitch rate. This state reconstructor uses the following equations:

$$\frac{dV}{dt} = - a_n \sin \alpha + a_x \cos \alpha$$

$$\frac{d\theta}{dt} = q \tag{8}$$

which have been derived from the definitions of a_n, a_x, and q (Fig. 1). Initial conditions for this reconstructor are obtained using additional measurements of airspeed and pitch angle.

Finite-Memory Identifier

The vehicle's stability derivatives are obtained by means of the linear model of the vehicle's short-period dynamics derived in Part II:

$$\Delta\dot{x} = A(p)\,\Delta x + b(p)\,\Delta\delta_e + c(p) \tag{9}$$

where $A(p)$, $b(p)$, and $c(p)$ may be determined from (3), and p is the unknown parameter vector:

$$p^T = (C_{L_\alpha}\ C_{m_\alpha}\ C_{m_q}\ C_{L_0}\ C_{m_o}\ C_{m_{\delta_e}}) \tag{10}$$

which is time varying, since both the stability derivatives and the off-time coefficients are functions of the vehicle's state.

The finite-memory identifier employed here assumes that p can be assumed to remain constant during some finite time arc length N. The parameter vector p is then determined on-line by seeking a least-squares fit between the model's response and a finite number N of stored measurements. The stored measurements are updated every time that a new set of measurements is taken by adding the new set of measurements and deleting the earliest set from the identifier's memory (Fig. 8). The procedure is, therefore, as follows:

Assuming that p is constant during the interval $(k_o, k_o + N - 1)$, with k_o some initial instant, an initial estimate $\hat{p}_{k_o}$ is obtained minimizing the following output error:

$$J(p) = \frac{1}{2} \sum_{k=k_o}^{k_o+N-1} \left\| z_{k+1} - x_{k+1} \right\|^2_{W_{k_o}} \tag{11}$$

subject to (9) with initial conditions

$$\Delta x_{k_o} = z_{k_o} - x_{nom,k_o} \tag{12}$$

where z_{k+1} are measurements of flight-path angle, pitch rate, and pitch angle, W_{k_o} is the following weighting matrix

$$W_{k_o} = \text{diag}\,(V_{nom,k_o}\ 1\ 1) \tag{13}$$

and x_{k+1}, the model's state at the $k + 1$ th instant, is given by

$$x_{k+1} = \Delta x_{k+1} + x_{nom,k_o} \tag{14}$$

Subsequent estimates $\hat{p}_{k_o+1}, \hat{p}_{k_o+2}, \ldots$, are then obtained in the same manner with k_o being replaced by $k_o + 1, k_o + 2, \ldots$ in (11), (12), (13), and (14).

A modified Newton-Raphson method has been employed to minimize $J(p)$ and obtain the new estimates $\hat{p}_{k_o+1}, \hat{p}_{k_o+2}, \ldots$ on-line. The iterative procedure is the following:

(i) Calculate $g(p)$ and $S(p)$ at $p = p_{k_o}$ as follows:

$$g(p) = \sum_{k=k_o}^{k_o+N-1} (z_{k+1} - x_{k+1})^T W_{k_o} \frac{\partial \Delta x_{k+1}}{\partial p} \tag{15}$$

$$S(p) = \sum_{k=k_o}^{k_o+N-1} \frac{\partial \Delta_{k+1}^T}{\partial p} W_{k_o} \frac{\partial \Delta x_{k+1}}{\partial p} \tag{16}$$

(ii) Determine the new estimate $\hat{p}_{k_o+1}$ from

$$\hat{p}_{k_o+1} = \hat{p}_{k_o} + S^{-1}(p_{k_o})\, g(p_{k_o}) \tag{17}$$

(iii) Update nominal conditions

(iv) Set $k_o = k_o + 1$ and repeat procedure.

The performance of this identifier is illustrated in Figure 9 which includes the reentry trajectory along which the parameters were identified. The length of the identifier's memory was limited to a total of 15 measurements taken each 0.125 sec. Furthermore, a square wave of 1^o amplitude and frequency 4 c/s was applied to the elevator to guarantee the identification of $C_{m_{\delta_e}}$. The results show clearly that the finite-memory identifier has good tracking capabilities, and updates its estimates as flight conditions change. Also, the C_{m_q} estimates indicate that the vehicle's response is not very sensitive to this parameter, as it was expected, due to the almost complete absence of damping on the pitch response.

Gain Computer

The gain computer's task consists of calculating the vehicle's flying qualities based on the identified stability derivatives, and determining whether the requirements specified in (6) are satisfied. If such requirements are not satisfied, the vehicle's flying qualities are modified by means of the following linear feedback law:

$$\delta_e = \delta_S + f_q\,(q - q_{nom}) + f_\theta\,(\theta - \theta_{nom}) \tag{18}$$

where δ_S is the stick deflection, and the feedback gains f_q and f_θ are calculated to yield acceptable damping and frequency characteristics. The feedback gains f_q and f_θ are determined by solving the linear system which results from equating the closed-loop characteristic polynomial with the desired characteristic polynomial.[6]

Adaptive Control System Results

A simulated flight of the reentry vehicle with the adaptive control system incorporated is included in Figure 8. The adaptive control system has corrected the very low damping existing on the original vehicle, yielding appropriate damping characteristics throughout the entire trajectory. The good natural frequency characteristics of the vehicle are reflected on the small values of the pitch-angle feedback gain. Also, the change of feedback gains indicates the decrease on $C_{m_{\delta_e}}$ as the vehicle slows down.

IV. CONCLUDING REMARKS

In this paper an adaptive control system which employs a finite-memory identifier has been introduced. Due to the finite memory of this identifier, modeling errors are kept finite. Furthermore, the design of a control for the vehicle is facilitated by the linear model. It is thus possible to derive a control law by using the updated linear model as an accurate representation of the vehicle.

V. BIBLIOGRAPHY

1. Anon.: "Flying Qualities of Piloted Airplanes." Military Specifications MIL-F-8785B (ASG), August 1969.

2. R. J. Fitzgerald, "Error Divergence in Optimal Filtering Problems." 2nd IFAC Symposium on Automatic Control, Austria, 1967.

3. F. M. Schlee, C. J. Standish, and N. F. Toda, "Divergence in the Kalman Filter." AIAA, Vol. 5, No. 6, June 1967, pp. 1114-1120.

4. A. H. Jazwinski, "Limited Memory Optimal Filtering." IEEE Trans. Automatic Control, Vol. AC-13, October 1968, pp. 558-563.

5. B. Etkins, Dynamics of Flight, New York. Wiley, 1959.

6. S. G. Gupta and L. Hasdorff, Fundamentals of Optimal Control, New York. Wiley, 1970.

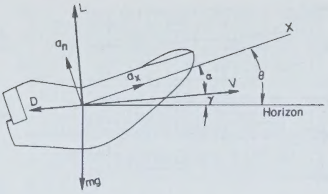

L
a_n
X
a_x
α
θ
V
D
γ
Horizon
mg

Figure 1.- Vehicle reference frame.

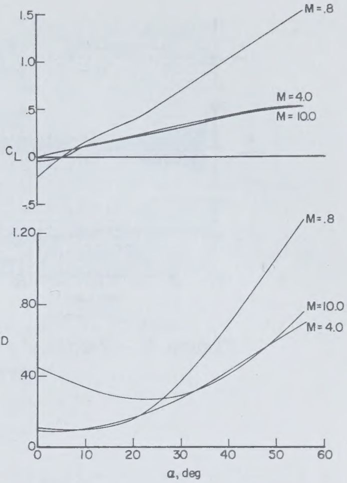


Figure 2.- Lift and drag coefficients.

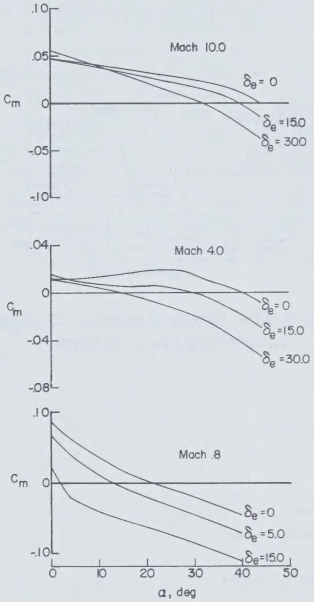


Figure 3.- Pitching-moment coefficient.

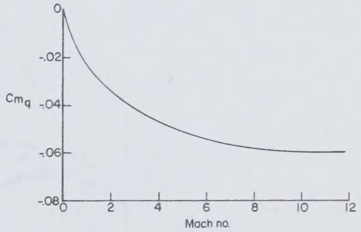


Figure 4.- Pitch-damping coefficient.

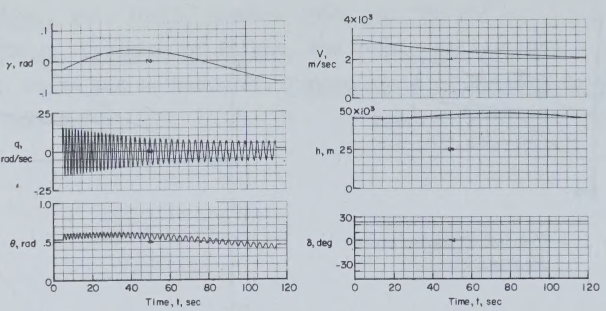


Figure 5.- Vehicle's response to a 5° off-trim elevator step.

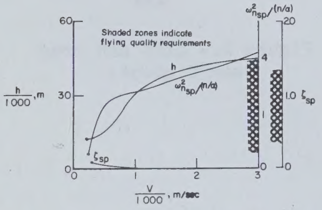
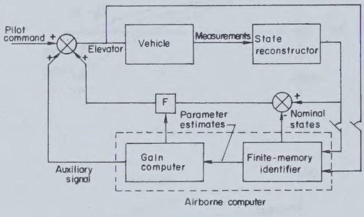
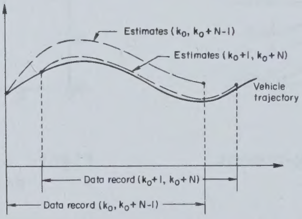


Figure 6.- Vehicle's short-period characteristics.



Pilot command
+
+
+
Elevator
Vehicle
Measurements
State reconstructor
F
Parameter estimates
+
-
Nominal states
Gain computer
Finite-memory identifier
Auxiliary signal
Airborne computer

Figure 7.- Block diagram of adaptive control system.



Estimates (k_0, k_0 + N-1)
Estimates (k_0+1, k_0+ N)
Vehicle trajectory
Data record (k_0+1, k_0+ N)
Data record (k_0, k_0+ N-1)

Figure 8.- Shifting process on a finite-memory identifier.

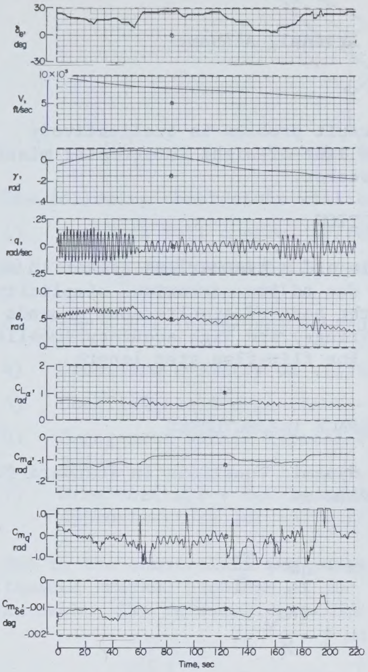


Figure 9.- Parameter identification in simulated flight.

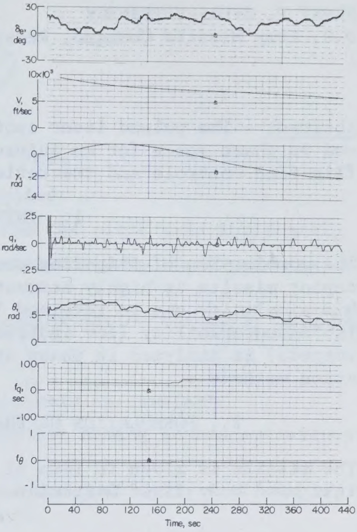


Figure 10.- Simulated flight with adaptive control system.

ON PULSED LASER SPOT HOMING MISSILES

ERNST T. EVERS-EUTERNECK
U.S. Army Missile Command, Redstone Arsenal, Alabama

Abstract - The pulsed laser spot pursuit problem is qualitatively analysed in phase space for an autonomous line-of-sight interceptor missile with flip-flop controls and trackerless seeker.

1. INTRODUCTION

The combination of flip-flop controls and a trackerless seeker in an interceptor missile is highly attractive due to its mechanical simplicity. How this simplicity can be retained for the case of simultaneous demands for high maneuverability (for indirect firings) and high accuracy (mobile spot targets) is analyzed in phase space for flip-flop step length modulation methods.

2. FORMULATION OF PHASE SPACE TRAJECTORIES

The missile is given in form of its dynamic equations in standard form for right hand body fixed Cartesian coordinates:

$$
\begin{aligned}
\dot{u} &= g_x - \frac{q'S}{M}\left[C_{\Delta 0} + C_{A\delta 2}\left(\delta_Q^2 + \delta_R^2\right)\right] + rv - qw + \frac{Th}{M} \\
\dot{v} &= g_y - \frac{q'S}{M}\left[C_{n\beta}\,\beta - C_{n\delta}\,\delta_r\right] + pw - ru + \frac{Th\epsilon_v}{M} \\
\dot{w} &= g_z - \frac{q'S}{M}\left[C_{n\alpha}\,\alpha - C_{n\delta}\,\delta_Q\right] + qu - pv + \frac{Th\epsilon_w}{M} \\
\dot{p} &= \frac{q'SD}{I_x}\left[C_{\ell\delta}\,\delta_Q\,\delta_R + \frac{D}{2V_M}\;C_{\ell P}\,p\right] + (\dot{r} + pq)\,\frac{I_{xz}}{I_x} \\
\dot{q} &= \frac{q'SD}{I_y}\left[C_{M\alpha}\alpha + C_{M\delta}\,\delta_Q + \frac{D}{2V_m}\left(C_{M\lambda}\,\dot{\alpha} + C_{M\sigma}\,q\right)\right] + \frac{I_{xz}}{I_y}\left(qr - p^2\right) \\
&\quad - \frac{I_x - I_z}{I_y}\,pr + \frac{ThL\epsilon_v}{I_y} \\
\dot{r} &= -\frac{\dot{q}SD}{I_z}\left[C_{M\beta}\,\beta + C_{M\delta}\,\delta_R + \frac{D}{2V_m}\left(C_{M\lambda}\,\beta - C_{M\sigma}\,r\right)\right] - \frac{I_{xz}}{I_z}\left(qr - p^2\right) \\
&\quad - \frac{I_y - I_x}{I_z}\,pq + \frac{ThL\epsilon_w}{I_z}
\end{aligned}
\tag{1}
$$

where δ is the wing deflection angle.

The coordinates are shown in Figure 1.

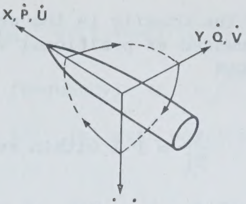


Figure 1

The following constraints will be implied:

1) $p = 0$ (A roll control system keeps the pitch axis horizontal)
2) $u = u_c$ (sustained boost compensates for drag after reaching nominal speed)
3) $\frac{\delta}{|\delta|} = \pm 1$ (bang-bang controls, approximation for quasi bang-bang)
4) Cruciform symmetry of airframe
5) Coefficients are constant - atmosphere is constant
6) Engagement within typical semi-active homing bounds and without noise
7) Wing controls - nearly torqueless and justifying small-angle approximations for angle of attack.

Due to the constraints the pursuit problem can be formulated as two simultaneous tracker problems of the line of sight in the planes of pitch and yaw each. The relevant equations are:

$$\begin{aligned}
\dot{v} &= -\frac{q'S}{M}\left[C_{N\beta}\beta - C_{N\delta}\delta_R\right] - u_c r \\
\dot{r} &= -\frac{q'SD}{I_z}\left[C_{M\beta}\beta + C_{M\delta}\delta_R + \frac{D}{2V_M}\left(C_{M\lambda}\dot{\beta} - C_{M\sigma} r\right)\right] \\
\dot{w} &= -\frac{qS}{M}\left[C_{N\alpha}\alpha - C_{N\delta}\delta_Q\right] + u_c q + g_w \\
\dot{q} &= \frac{q'SD}{I_y}\left[C_{M\alpha}\alpha + C_{M\delta}\delta_Q + \frac{D}{2V_M}\left(C_{M\lambda}\dot{\alpha} + C_{M\sigma} q\right)\right].
\end{aligned} \qquad (2)$$

Each planer motion can be investigated independently. The yaw-plane motions will be studied first.

Small angle approximations are introduced: $\beta = \frac{v}{u_c}$; $\dot{\beta} = \frac{\dot{v}}{u_c}$.

The coefficients for the yaw equations are expressed in more compact notation with these definitions:

$$C_{vv} = \frac{q'S}{Mu_c} C_{N\beta} \qquad C_{v\delta} = \frac{q'S}{M} C_{N\delta} \qquad C_{vr} = u_c$$

$$C_{rr} = \frac{q'SD^2}{2I_y u_c} C_{M\sigma} \qquad C_{r\delta} = \frac{q'SD}{I_y} C_{M\delta} \qquad C_{rv} = \frac{q'SD}{I_y u_c} C_{M\beta} .$$

Since the misalignment angle of the missile is the observed system error, the dependent variables are expressed as positional variables: $\int \dot{v}dt = V(t)$; $\int \dot{r}dt = \dot{R}(t)$. Equations (2) become

$$\begin{aligned} &\ddot{V} + C_{vv}\dot{V} + C_{vr}\dot{R} + C_{v\dot{r}}\ddot{R} = C_{v\delta}\delta \\ &\ddot{R} + C_{rr}\dot{R} - C_{rv}\dot{V} - C_{r\dot{v}}\ddot{V} = C_{r\delta}\delta \end{aligned} \qquad \frac{\delta}{|\delta|} = \pm\ 1 \text{ within switching intervals.} \tag{3}$$

The term $- C_{r\dot{v}}\ddot{V}$ is synthetic and will facilitate the analysis later. For the purpose of analytical decoupling of V and R Laplace transforms are used. The Laplace transforms are defined: $L\left\{V(t)\right\} = \varphi(s)$; $L\left\{R(t)\right\} = \rho(s)$. In the s-domain Equations (3) become

$$\begin{aligned} s\left(s + C_{vv}\right)\varphi + s\left(sC_{v\dot{r}} + C_{vr}\right)\rho &= \frac{1}{s}C_{v\delta}\delta + s\left(V(0) + C_{v\dot{r}}R(0)\right) + \dot{V}(0) \\ &\quad + C_{vv}V(0) + C_{vr}R(0) + C_{v\dot{r}}\dot{R}(0) \\ -s\left(sC_{r\dot{v}} + C_{rv}\right)\varphi + s\left(s + C_{rr}\right)\rho &= \frac{1}{s}C_{r\delta}\delta + s\left(-C_{r\dot{v}}V(0) + R(0)\right) - C_{r\dot{v}}\dot{V}(0) \\ &\quad - C_{rv}V(0) + C_{rr}R(0) + \dot{R}(0)\,. \end{aligned} \tag{4}$$

One can extract from Equations (4) directly:

a) The system's determinant $D = \begin{vmatrix} s^2 + sC_{vv} & s^2C_{v\dot{r}} + sC_{vr} \\ -s^2C_{r\dot{v}} - sC_{rv} & s^2 + sC_{rr} \end{vmatrix}$

$$= \left(1 + C_{r\dot{v}}C_{v\dot{r}}\right)s^2\left[s^2 + \frac{s\left(C_{rr} + C_{vv} + C_{r\dot{v}}C_{vr} + C_{v\dot{r}}C_{rv}\right) + C_{vv}C_{rr} + C_{rv}C_{vr}}{1 + C_{r\dot{v}}C_{v\dot{r}}}\right]$$

$$= As^2\left[s^2 + 2s\zeta\ \omega_0 + \omega_0{}^2\right]$$

b) The system function $C = s^2 + 2s\zeta\omega_0 + \omega_0{}^2 = s + \zeta\omega_0{}^2 + \omega_n{}^2\omega_0{}^2$

c) The characteristic equation $s^2 + 2s\zeta\omega_0 + \omega_0{}^2 \equiv 0$.

Note: In practices the constant $A \equiv 1 + C_{r\dot{v}}C_{v\dot{r}} \approx 1$.

The following standard definitions are implied:

a) Natural frequency of the undamped system: ω_0; where

$$\omega_0{}^2 = \frac{1}{A}\left(C_{vv}C_{rr} + C_{rv}C_{vr}\right)$$

b) Relative damping coefficient: $\zeta = \frac{1}{2A\omega_0}\left(C_{rr} + C_{vv} + C_{vr}C_{r\dot{v}} + C_{rv}C_{v\dot{r}}\right)$.

c) Damped natural frequency: $\omega_n = \sqrt{1 - \zeta}$

Equations (4) are solved for $\varphi(s)$ with positive deflection angle δ

$$\varphi(s) = \frac{1}{D}\left[D_{\varphi,\delta}\frac{\delta}{s} + D_{\varphi,v}V(0) + D_{\varphi,R}R(0) + D_{\varphi,\dot{v}}\dot{V}(0) + D_{\varphi,\dot{R}}\dot{R}(0)\right]$$

where

$$D_{\varphi,\delta} = \begin{vmatrix} C_{v\delta} & s\left(sC_{v\dot{r}} + C_{vr}\right) \\ C_{r\delta} & s\ s + C_{rr} \end{vmatrix} = \begin{cases} = s^2\left(C_{v\delta} - C_{r\delta}C_{v\dot{r}}\right) + s\left(C_{v\delta}C_{rr} - C_{r\delta}C_{vr}\right) \\ = s^2F'_{v1} + sF'_{v2} \end{cases}$$

$$D_{\varphi,v} = \begin{vmatrix} s + C_{vv} & s\left(sC_{v\dot{r}} + C_{vr}\right) \\ -sC_{r\dot{v}} - C_{rv} & s\left(s + C_{rr}\right) \end{vmatrix} = \frac{D}{s}$$

$$D_{\varphi,R} = \begin{vmatrix} sC_{v\dot{r}} + C_{vr} & s\left(sC_{v\dot{r}} + C_{vr}\right) \\ s + C_{rr} & s\left(s + C_{rr}\right) \end{vmatrix} = 0$$

$$D_{\varphi,\dot{v}} = \begin{vmatrix} 1 & s\left(sC_{v\dot{r}} + C_{vr}\right) \\ -C_{r\dot{v}} & s\left(s + C_{rr}\right) \end{vmatrix} = \begin{cases} = s^2\left(1 + C_{r\dot{v}}C_{v\dot{r}}\right) + s(C_{rr} + C_{r\dot{v}}C_{vr} \\ = s^2A + sM'_v \end{cases}$$

$$D_{\varphi,\dot{R}} = \begin{vmatrix} C_{v\dot{r}} & s\left(sC_{v\dot{r}} + C_{vr}\right) \\ 1 & s\left(s + C_{rr}\right) \end{vmatrix} = s\left(C_{v\dot{r}}C_{rr} - C_{vr}\right) = sN'_v \ .$$

The solution for $\rho(s)$ has the same form; coefficients are obtained by consistent exchange of indices v and r

$$\varphi(s) = \frac{F'_{v1}\delta}{AsC} + \frac{F'_{v2}\delta}{As^2C} + \frac{V(0)}{s} + \frac{\dot{V}(0)}{C} + \frac{M'_v\dot{V}(0)}{AsC} + \frac{N'_v\dot{R}(0)}{AsC}$$

$$\rho(s) = \frac{F'_{R1}\delta}{AsC} + \frac{F'_{R2}\delta}{As^2C} + \frac{R(0)}{s} + \frac{\dot{R}(0)}{C} + \frac{M'_R\dot{R}(0)}{AsC} + \frac{N'_R\dot{V}(0)}{AsC} \ . \tag{5}$$

The time-domain solution $L^{-1}\left\{\varphi(s)\right\} = V(t)$ and $L^{-1}\left\{\rho(s)\right\} = R(t)$ are also of identical analytical structure due to the synthetic terms inserted in Equations (3). It therefore suffices to write down only one of them. The other one is obtained by consistent swap of coefficient indices and the dependent variable as will be shown.

For V(t) one obtains:

$$V(t) = \frac{F'_{v2}\delta}{\omega_0{}^3 A}\ \omega_0 t + V(0) + \frac{1}{\omega_n}\,\mathrm{Im}\left\{\left[\left(\frac{F_{v1}}{A\omega_0{}^2} + \frac{F'_{v2}}{A\omega_0{}^3}\,e^{-j\psi}\right)\delta + \frac{M'_v\dot{V}(0) + N'_v\dot{R}(0)}{A\omega_0{}^2}\right]\right.$$
$$\left.\left[e^{-j\psi}\left(-1 + e^{(j\omega_n - \zeta)\omega_0 t}\right)\right] + \frac{1}{\omega_0}\,\dot{V}(0)e^{(j\omega_n - \zeta)\omega_0 t}\right\}$$

where $\psi = \arctan \frac{\omega_n}{-\zeta}$. For shorter notation coefficients are redefined as follows:

$F_{v1} = \frac{F'_{v1}}{A\omega_0{}^2}$; $F_{v2} = \frac{F'_{v2}}{A\omega_0{}^3}$; $M_v = \frac{M'_v}{A\omega_0{}^2}$; $N_v = \frac{N'_v}{A\omega_0{}^2}$; $M = \frac{1}{\omega_0}$; and analoguously

for coefficients in R(t). By use of the new coefficient notation and expedient scaling of $\dot{V}(t)$ it can be shown that phase plane trajectories in the $(V, \dot{V})$ plane have simplest expressions by defining the phase point locus in the complex plane for an oblique coordinate system: $\hat{v}(t) = V(t) - \frac{e^{-j\psi}}{\omega_0}\cdot\dot{V}(t)$. With newly defined coefficients, one obtains:

$$V(t) = F_{v2}\omega_0 t\delta + \quad V(0) + \frac{1}{\omega_n}\,\mathrm{Im}\left\{\left[\left(F_{v1} + F_{v2}e^{-j\psi}\right)\delta + M_v\dot{V}(0) + N_v\dot{R}(0)\right]\right.$$
$$\left.\left[e^{-j\psi}\left(-1 + e^{(j\omega_n - \zeta)\omega_0 t}\right)\right] + M\dot{V}(0)e^{(j\omega_n - \zeta)\,\omega_0 t}\right\}$$

$$\frac{\dot{V}(t)}{\omega_0} = \frac{1}{\omega_n}\,\mathrm{Im}\left\{\left(F_{v1} + F_{v2}e^{-j\psi}\right)\delta\left(-1 + e^{(j\omega_n - \zeta)\,\omega_0 t}\right) + \left(M_v\dot{V}(0)\right.\right.$$
$$\left.\left. + N_v\dot{R}(0)\right)e^{(j\omega_n - \zeta)\,\omega_0 t} + Me^{j\psi}\dot{V}(0)e^{(j\omega_n - \zeta)\omega_0 t}\right\}.$$

The phase point locus is

$$\hat{v}\ (t) = V(0) + F_{v2}\omega_0 t\,\delta + \left[M_v\dot{V}(0) + N_u\dot{R}(0) + \left(F_{v1} + F_{v2}e^{j\psi}\right)\delta\right]$$
$$\left[1 - e^{-(j\omega_n + \zeta)\omega_0 t}\right] - e^{-j\psi}M\dot{V}(0)e^{-(j\omega_n + \zeta)\,\omega_0 t}\ .$$

For further simplification more reductions are made:

a) If $F_{v1} + F_{v2}e^{j\psi} = \hat{F}_v$, one has $|\hat{F}_v| = F_v = \left[\left(F_{v1} - \zeta F_{v2}\right)^2 + \omega_n^{\,2} F_{v2}^2\right]^{\frac{1}{2}}$

$= \left[F_{v1}^2 - 2\zeta F_{v1}F_{v2} + F_{v2}^2\right]^{\frac{1}{2}}$ and $\hat{F}_v = F_v e^{j\kappa_v}$ with $\tan \kappa_v = \dfrac{\omega_n F_{v2}}{F_{v1} - \zeta F_{v2}}$

b) The phase point locus is normalized by using $\dfrac{\hat{v}}{F_v} = \hat{p}_v$ which makes $\dfrac{\hat{F}_v}{F_v} = e^{j\kappa_v}$

c) Considering that the phase point loci $\hat{v}(t) = V(t) - e^{-j\psi}\dfrac{\dot{V}(t)}{\omega_0}$ and $\hat{r}(t) = R(t) - e^{-j\psi}\dfrac{\dot{R}(t)}{\omega_0}$ are intercoupled by the terms $N_v\dot{R}(0)$ and $N_R\dot{V}(0)$, and considering a and b, the following redefinitions for N_v and N_R will allow direct insertion of reduced coupling terms in the normalized phase point loci.

$\dfrac{N_v}{F_v}\dot{R}(0)$ becomes $\dfrac{N_v}{F_v} \cdot F_R \cdot \dfrac{\dot{R}(0)}{F_R} = P_v \cdot \dot{\epsilon}_R(0)$ with $P_v = \dfrac{N_v}{F_v} F_R$ and

$\dfrac{N_R}{F_R}\dot{V}(0)$ becomes $\dfrac{N_R}{F_R} \cdot F_v \cdot \dfrac{\dot{V}(0)}{F_v} = P_R \cdot \dot{\epsilon}_v(0)$ with $P_R = \dfrac{N_R}{F_R} \cdot F_v$.

The reduced phase point loci for positive δ are:

$$\hat{p}_v = \frac{\hat{v}}{F_v} = \frac{1}{F_v}\left[V(t) - e^{-j\psi}\frac{\dot{V}(t)}{\omega_0}\right] =$$

$$= \epsilon_v(0) + \frac{F_{v2}}{F_v}\omega_0 t\,\delta + e^{-j\psi}\left[M_v\dot{\epsilon}_v(0) + P_v\dot{\epsilon}_R(0) + e^{j\kappa_v}\delta\right] \quad (6v)$$

$$\left[1 - e^{-\left(j\omega_n + \zeta\,\omega_0 t\right.}\right] - e^{-j\psi}M\dot{\epsilon}_v(0)e^{-\left(j\omega_n + \zeta\right)\omega_0 t}$$

$$\hat{p}_r = \frac{\hat{r}}{F_R} = \frac{1}{F_R}\left[R(t) - e^{-j\psi}\frac{\dot{R}(t)}{\omega_0}\right] =$$

$$= \epsilon_R(0) + \frac{F_{R2}}{F_R}\omega_0 t\,\delta + e^{-j\psi}\left[M_R\dot{\epsilon}_R(0) + P_R\dot{\epsilon}_v(0) + e^{-j\kappa_R}\delta\right] \quad (6r)$$

$$\left[1 - e^{-\left(j\omega_n + \zeta\right)\omega_0 t}\right] - e^{-j\psi}M\dot{\epsilon}_R(0)e^{-\left(j\omega_n + \zeta\right)\omega_0 t} .$$

The phase point locus (or phase trajectory) will serve as the analytical tool for the development of the guidance loop. A typical phase trajectory is shown in Figure 2 for $\hat{p}_r$, the angular system with positive δ. To reveal the topological structure of the typical phase trajectory, $\hat{p}_r$ is reformulated as follows:

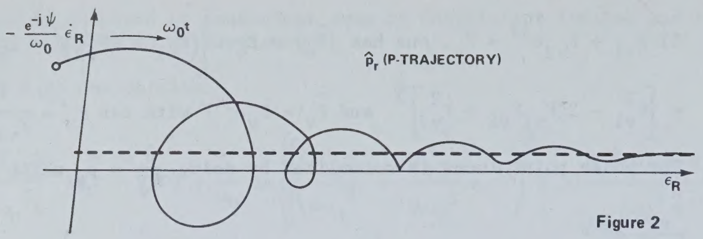

$-\frac{e^{-j\psi}}{\omega_0}\dot{\epsilon}_R$
$\omega_0 t$
$\hat{p}_r$ (P-TRAJECTORY)
ϵ_R

Figure 2

$$\hat{p}_r = \left\{ \begin{array}{c|c} \epsilon_r(0) + \left[M_r\dot{\epsilon}_r(0) + P_r\dot{\epsilon}_v(0)\right] & + e^{j\kappa r_\delta} \quad + \frac{F_{r2}}{F_r}\,\omega_0 t\delta + \\ \hline -\left[M_r\dot{\epsilon}_r(0) + P_r\dot{\epsilon}_v(0)\right]\, e^{-(j\omega_n + \zeta)\omega_0 t} & - e^{j\kappa r_\delta} e^{-(j\omega_n + \zeta)\omega_0 t} \\ -e^{-j\psi}M\dot{\epsilon}_r(0)e^{-(j\omega_n + \zeta)\omega_0 t} & \end{array} \right\} \qquad (7)$$

The right member is arranged to fill four quadrants. The following is revealed:

a) The lower quadrants contain expressions for a logarithmic spiral (the generatrix)

b) The upper quadrants contain expressions for the locus of the focal point of the generatrix

c) The left quadrants contain the natural solution

d) The right quadrants contain the forced solution

e) The upper right quadrant describes the maneuverability of the missile

f) $\hat{p}_r$ describes a trochoid about the line $e^{j\kappa r_\delta} + \frac{F_{r2}}{F_r}\omega_0 t\delta$ with attenuation factor - ζ (Figure 2)

g) Equation (7) represents a P - trajectory for position δ and an N-trajectory for negative δ

h) Time is proportional to focal angles on the generatrix (this is due to the choice of oblique phase plane coordinates)

According to Equation (3), the pursuit loop is now formed (Figure 3) and equipped with an exchangeable compensation branch. The loop is formed for one planar control, e.g., yaw.

3. THE ALIGNMENT PROBLEM

Upon acquisition the full-helm alignment is initiated. This phase ends with the first commutated seeker output (Figure 3a).

4. THE REGULATOR PROBLEM

To eliminate large size and irregular oscillations of alternating P- and N-trajectories about this target point as would take place with simple cross-hair switchings, the following method is suggested:

The relation between delays, switching intervals, and corresponding steady state limit cycles is now developed. Let fixed delays be much shorter than a chosen limit cycle period $\omega_0 T$. Let $\frac{1}{2}\,\omega_0 T = \tau$. Assume a limit cycle symmetric to the origin in the phase plane. Assign the beginning of a switch interval to $\omega_0 T = 0$ and switch point PN, and the end of this switch interval to $\omega_0 t = \tau$ and switch point NP.

In the phase plane one obtains by use of symmetry conditions:

$$\epsilon_R(0) - e^{-j\psi}\,\frac{\dot{\epsilon}_r(0)}{\omega_0} = -\epsilon_R(\tau) + e^{-j\psi}\,\frac{\dot{\epsilon}_R(\tau)}{\omega_0}\ .$$

Equation (7) is used for a cyclic half interval between the symmetric switch points $\hat{p}_R(0)$ and $\hat{p}_R(\tau)$ with $\frac{\delta}{|\delta|} = -1$ and $I_{R0} = M_r\,\dot{\epsilon}_R(0) + P_r\,\dot{\epsilon}_v(0)$. One obtains by observing cyclic symmetry conditions:

$$\begin{aligned}\hat{p}_R(t) = {} & \epsilon_R(0) + I_{R0} \qquad - e^{j\kappa R_\delta} \qquad - \frac{r_{R2}}{F_R}\,\delta\tau \\ & - I_{R0}e^{-(j\omega_n + \zeta)\tau} - e^{-j\psi}M\dot{\epsilon}_R(0)e^{-(j\omega_n + \zeta)\tau} + e^{j\kappa R_\delta}e^{-(j\omega_n + \zeta)\tau}\end{aligned} \tag{8}$$

Assuming now $\hat{p}_R(\tau)$ in position and formulating Equation (7) for no elapsing time one obtains:

$$\begin{aligned}\hat{p}_R(\tau) = {} & -\epsilon_R(0) - I_{R0} \qquad - e^{j\kappa R_\delta} \\ & + I_{R0} \qquad + e^{-j\psi}M\dot{\epsilon}_R(0) \qquad + e^{j\kappa R_\delta}\ .\end{aligned} \tag{9}$$

Summing Equations (8) and (9) results in

$$2\hat{p}_R(\tau) = \left[e^{-j\psi}M\dot{\epsilon}_R(0) + I_{R0} - e^{j\kappa R_\delta}\right]\left[1 - e^{-(j\omega_n + \zeta)\tau}\right] - \frac{F_{R2}}{F_R}\,\tau\,\delta\ . \tag{10}$$

The difference in the spiral loci of Equations (8) and (9) must vanish:

$$0 = 2\,\epsilon_R(0) + 2I_{R0} - \frac{F_{R2}}{F_R}\,\tau\,\delta\ . \tag{11}$$

The difference in the focal loci of Equations (8) and (9) also must vanish:

$$0 = \left[e^{-j\psi}M\dot{\epsilon}_R(0) + I_{R0}\right]\left[1 + e^{-(j\omega_n + \zeta)\tau}\right] + e^{j\kappa R_\delta}\left[1 - e^{-(j\omega_n + \zeta)\tau}\right].$$

This can be written as

$$\left(Me^{-j\psi} + M_v\right)\dot{\epsilon}_R(0) + P_R\,\dot{\epsilon}_V(0) = -e^{j\kappa R_\delta}\cdot\frac{1 - e^{-(j\omega_n + \zeta)\tau}}{1 + e^{-(j\omega_n + \zeta)\tau}}\ . \tag{12}$$

Introducing shorter notation: $\hat{Q}_v = M_v + Me^{-j\psi}$ and $\hat{H}(\tau) = \dfrac{1 - e^{-\,j\omega_n + \zeta\,\tau}}{1 + e^{-\,j\omega_n + \zeta\,\tau}}$.

Equation (12), applied for the angular and lateral system is used to find cyclic terms for $\dot{\epsilon}_R$ and $\dot{\epsilon}_v$ as function of $H(\tau)$ representing τ. The simultaneous Equations (12) are

$$\begin{aligned} \hat{Q}_v \dot{\epsilon}_v(0) + P_v \dot{\epsilon}_R(0) &= - e^{j\kappa_v}\delta \, \hat{H}(\tau) \\ P_R \dot{\epsilon}_v(0) + \hat{Q}_R \dot{\epsilon}_R(0) &= - e^{j\kappa_R}\delta \, \hat{H}(\tau) \, . \end{aligned} \tag{13}$$

For cyclic switch points one obtains from Equation (13):

$$\dot{\epsilon}_{Rc}(0) = \frac{P_R e^{j\kappa_v} - \hat{Q}_v e^{j\kappa_r}}{\hat{Q}_v \hat{Q}_R - P_v P_R} \hat{H}(\tau)\delta \quad \text{and} \quad \dot{\epsilon}_{vc}(0) = \frac{P_v e^{j\kappa_R} - \hat{Q}_R e^{j\kappa_v}}{\hat{Q}_v \hat{Q}_R - P_v P_R} \hat{H}(\tau)\delta \quad .$$

For the solution for ϵ_{Rc} representing ϵ_{vc} the terms I_{Rc} representing I_{vc} are explicitly developed by insertion of the solutions for $\dot{\epsilon}_{Rc}(0)$ and $\dot{\epsilon}_{vc}(0)$:

$$I_{vc}(0) = M_v \dot{\epsilon}_{vc}(0) + P_v \dot{\epsilon}_{Rc}(0) = \frac{M_v\left(P_v e^{j\kappa_R} - \hat{Q}_R e^{j\kappa_v}\right) + P_v\left(P_R e^{j\kappa_v} - \hat{Q}_v e^{j\kappa_R}\right)}{\hat{Q}_v \hat{Q}_r - P_v P_r} \hat{H}(\tau)\,\delta \; .$$

From Equation (11) one obtains: $\epsilon_{vc}(0) = \frac{F_{v2}}{F_v} \cdot \frac{\tau}{2}\delta - I_{vc}(0)$. The cycle switch point PN for a given τ is now expressed for the lateral system:

$$\begin{aligned} \hat{p}_{vc}(\tau)_0 &= \epsilon_{vc}(0) - e^{-j\psi} \frac{\dot{\epsilon}_{vc}(0)}{\omega_0} \\ &= \frac{F_{v2}}{F_v} \cdot \frac{\tau}{2}\,\delta - \frac{M_v\left(P_v\, e^{j\kappa_R} - \hat{Q}_R e^{j\kappa_v}\right) + P_v\left(P_R e^{j\kappa_R} - \hat{Q}_v e^{j\kappa_R}\right)}{\hat{Q}_v \hat{Q}_R - P_v P_R} \\ &\qquad \frac{+ Me^{-j\psi}\left(P_v e^{j\kappa_R} - \hat{Q}_R e^{j\kappa_v}\right)}{} \hat{H}(\tau)\,\delta \\ &= \frac{F_{v2}}{F_v} \cdot \frac{\tau}{2}\,\delta + e^{j\kappa_v} \hat{H}(\tau)\,\delta \end{aligned}$$

$$\hat{p}_{vc}(\tau)_0 = \left[\frac{F_{v2}}{F_v} \cdot \frac{\tau}{2} + e^{j\kappa_v}\hat{H}(\tau)\right]\delta$$

and, analogously $\hat{p}_{vc}(\tau)_0 = \left[\frac{F_{R2}}{F_r} \cdot \frac{\tau}{2} + e^{j\kappa_R}\hat{H}(\tau)\right]\delta$. For computational purposes $\hat{H}(\tau) = \frac{1 - e^{-(j\omega_n + \zeta)\tau}}{1 + e^{-(j\omega_n + \zeta)\tau}} = \frac{\mathrm{sh}\zeta\tau + j \sin \omega_n\tau}{\mathrm{ch}\zeta\tau + \cos \omega_n\tau}$.

Now if the first commutation of the seeker output is used to initiate control commutation with each laser pulse, regardless of the sign of the seeker outputs, P-and N-trajectories of length T_ℓ (laser pulse repetition period) will alternate and, after some transients, settle symmetrically in the phase plane abscissa (principle of least action). Symmetry with regard

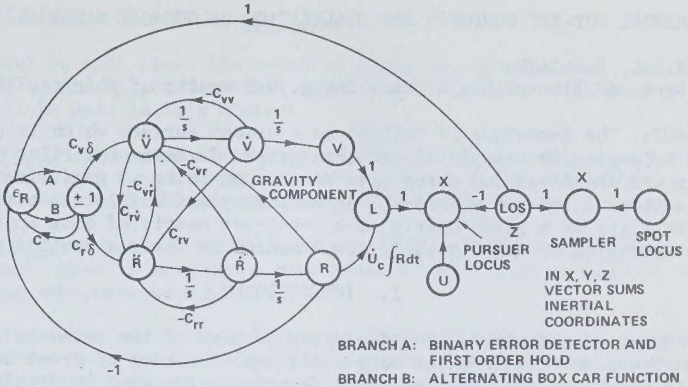


Figure 3

VERSION (3a) : WITHOUT BRANCHES B & C
VERSION (3b) : WITHOUT BRANCHES A & C
VERSION (3c) : WITHOUT BRANCH A

BRANCH A: BINARY ERROR DETECTOR AND FIRST ORDER HOLD
BRANCH B: ALTERNATING BOX CAR FUNCTION MODULATOR
BRANCH C: SEEKER SWIVEL LOOP, OR LINE OF SIGHT ARRESTING LOOP

to the target point is achieved by (1) imbedding the optical axis within a strip of angular width $\pm F_{r} R_{e} \left\{ \hat{p}_{R}\left(T_{\ell}\right) \exp^{j\left(\psi - \frac{\pi}{2}\right)} \right\}$ with T_{ℓ} the laser pulse repetition period [Equation (14)] and (2) modulation of the synchronous control commutations in such a way that the strip edges see pulses alternatively. The seeker outputs monitor the modulation through a Schmidt Trigger (Figure 3b).

5. THE FOLLOW-UP PROBLEM

If the regulator version (Figure 3b) cannot secure steady state symmetric locations of signals or seeker strip edges, the line of sight is rotated by target motion, the influence of gravity and wind. Assuming the target spot is not accelerating, it is only necessary to change the average course slightly faster than the line of sight (by slight modulation) and swing the seeker axis back at a lower rate (in missile coordinates). Empirically it was found that this method handles envisioned engagements with a fixed modulation and fixed loop swivel gain (Figure 3c) and that constant bearing pursuit is closely approximated.

6. CONCLUSIONS

The pursuit problem is solved for simple hardware and a pulsed laser spot target by use of standard AC circuit methods. Chronological adaptation of the compensation function to different stages of the pursuit game provide flexibility as needed for large initial conditions.

7. REFERENCES

J. Flügge-Lotz, Discontinuous Automatic Control, Princeton University, Princeton, New Jersey, 1953.

Gille, Pellegrin, Decaulne, Feedback Control Systems, McGraw-Hill Book Company, New York 1959.

M. Minorsky, Theory of Nonlinear Control Systems, McGraw-Hill Book Company, New York 1969.

FUNDAMENTAL CUT-SET CONTOURS AND REALIZATION OF CUT-SET MATRICES*

M.LAL, N.C.JAIN
Dept. of Electronics & Comm. Engrg.,University of Roorkee,INDIA.

ABSTRACT: The f-contour is defined as a closed surface which is pierced by edges belonging to some f-cut-set in a graph. Certain properties of f-contours are discussed which have been found to be of great importance in realization of cut-set matrices. The necessary and sufficient conditions for realizability of a given matrix as a f-cut-set matrix of some appropriate graph (directed or non-directed) are embodied in the realization technique.

I. INTRODUCTION

The concept of f-contours,though similar to that of the cut-sets corresponding to branches of a tree,has many additional features of great interest. In this paper,the characteristics of f-contours are used to develop a realization technique for f-cut-set matrices. However,the f-contours have also found applications in network analysis,and synthesis of short-circuit conductance matrices.

Some definitions regarding the f-contours are given in the next section,and a number of important theorems are presented in Section III. Finally,the proposed technique for realization of a f-cut-set matrix is developed in Section IV.

II. DEFINITIONS

DEFINITION 1. The k-th f-cut-set contour (or simply the k-th contour), denoted by Q_k,corresponding to the k-th branch of some tree is defined as a closed surface drawn such that it cuts those,and only those,edges of the graph which belong to the k-th f-cut-set. Also,it does not pass through any vertex of the graph.

It follows that a f-contour must be a 3-dimensional closed surface for non-planar graph;but in case of planar graphs a f-contour may also be represented by a plane closed curve. An edge is defined to "cut" a closed surface if it pierces the surface either once or any odd number of times. For oriented cut-sets an orientation,as defined below,is also assigned to each of the f-contours.

DEFINITION 2. The f-contour Q_k is defined to be oriented "outwards" ("inwards") if the orientation of the branch k of the relevant tree is from inside to outside (outside to inside) of the contour Q_k.

DEFINITION 3. The orientation of a contour Q_k with respect to an edge p,which cuts Q_k,is defined to be similar if their orientation arrows are both pointing wither outwards or inwards the contour Q_k;the Q_k and p are oriented oppositely if either of their orientation arrows is pointing outwards and the other inwards.

DEFINITION 4. An edge cutting a f-contour Q_k will be called "an edge of Q_k". The symbol Θ_k will be used to designate the set of all the edges of Q_k. Similarly,the set of all the edges common to all the contours $Q_p,Q_q,\ldots,$

---

* Condensed from a part of the Doctoral Dissertation titled "Realization of Resistive n-Port Networks,Cut-Set Matrices,and Loop Matrices",submitted by the junior author to University of Roorkee in 1968.

Q_t will be represented by the set $\theta_{p,q,\ldots,t}$.

It is meaningful to talk about the mutual orientation of a pair of f-contours, say Q_p and Q_q, provided that $\theta_{p,q}$ is not null. The following definition will be used in this context.

DEFINITION 5. The mutual orientation of Q_p and Q_q is defined to be positive (negative) if $\theta_{p,q}$ is not null and if Q_p and Q_q have similar (opposite) orientations with respect to any of the edges of the set $\theta_{p,q}$.

DEFINITION 6. Any two f-contours of a particular graph are defined to "intersect" each other if they cross each other at more than one points or, in the limiting case, meet at a single point.

Some important results can be established by distinguishing the space lying inside a f-contour from the space lying outside it. The following definitions are useful in this connection.

DEFINITION 7. The space enclosed by a f-contour Q_k will be called the "region inside Q_k" and it will be denoted by the symbol R_k. Similarly, the space lying outside Q_k will be called the "region outside Q_k" and it will be represented by the symbol R_k'.

Another region r_k which is different from R_k only when Q_k encloses another f-contour, is defined below.

DEFINITION 8. When a number of f-contours for a graph are drawn in a non-intersecting manner, the "region corresponding to a f-contour Q_k" is defined as the region which lies internal to Q_k but external to any other f-contour which may itself be enclosed by Q_k. The region corresponding to Q_k will be denoted by the symbol r_k. Any point, or any vertex of the graph, lying in the region r_k will be said to correspond to the f-contour Q_k.

Yet another region, namely the region between two contours, is defined next.

DEFINITION 9. A point t is defined to be "between" the f-contours Q_p and Q_q if it is possible to draw a line (not necessarily straight) from point t to any point on Q_p without cutting Q_q, and a line from point t to any point on Q_q without cutting Q_p. The "region between Q_p and Q_q" is defined to be the set of all points lying between Q_p and Q_q.

III. PROPERTIES OF f-CONTOURS

The uniqueness of mutual orientation of a pair of f-contours is stated below and it can be proved quite easily.

PROPERTY 1. If in a graph the f-contours Q_p and Q_q are such that the set $\theta_{p,q}$ is not null, and if Q_p and Q_q are oriented similarly (oppositely) through any one of the edges in the set $\theta_{p,q}$, the two f-contours must be oriented in this fashion through each of the remaining edges in $\theta_{p,q}$.

The following theorem is an important result concerning the f-contours.

THEOREM 1. All the f-contours corresponding to any tree T selected in a given graph (directed or non-directed) may always be so drawn that no two of the f-contours are intersecting.

PROOF: The theorem is proved below by presenting two standard procedures which will always yield all the f-contours for the tree T, without any one of them intersecting any of the others.

METHOD 1. We recall that a "tip branch" of a tree is an end branch of the tree such that one of the vertices on which it is incident has degree one as far as the tree is concerned. Let the set of the tip branches of T be represented by S_{t1}. The f-contours for all the branches of S_{t1} are drawn such that each encloses that particular vertex of the respective branch at which no other branch of T is incident. Clearly, all these contours can be drawn in a non-intersecting manner.

If all the branches of S_{t1} are removed from T, there would remain a connected subgraph of T, say T_1, whose tip branches will be represented by the set S_{t2}. At least one vertex of each of the branches of S_{t2} will have branches incident upon it only from the set S_{t1}. The f-contours are now drawn for branches of S_{t2}, enclosing in each case this particular kind of vertex. Obviously, no two of the contours drawn so far need intersect each other. Proceeding in this manner the theorem is proved quite easily.

METHOD 2. Let us select any of the tip branches of T, say i, and designate its vertices by v_i and v_i'; the latter being the vertex at which no other branch of T is incident. The contour Q_i is drawn enclosing the vertex v_i.

Let tree branches j,k,l,m,... with vertices (v_j,v_j'), (v_k,v_k'), (v_l,v_l'), (v_m,v_m'),..., respectively, be incident - apart from the branch i - on the vertex v_i. If branch j is removed from T, two separate subgraphs of T would be produced and the one containing the vertex v_j will be denoted by S_j. The f-contour corresponding to branch j is to be so drawn that it encloses S_j. Proceeding in this manner the theorem is again proved quite easily.

THEOREM 2. If branches i and j are selected arbitrarily in any tree T, and f-contours Q_i and Q_j are drawn so that neither of these contours encloses the other, then the rest of the f-contours for T can always be drawn in such a manner that all the f-contours are of non-intersecting nature.

PROOF: Let (v_i,v_i') and (v_j,v_j') be the vertices of the branches i and j, respectively; such that v_i and v_j are the terminal vertices of a path in T which has i and j as its terminal branches. Note that Q_i must enclose v_i, and Q_j must enclose v_j. The three subgraphs of T which would be obtained when branches i and j are both removed from T and which contain vertices v_i, v_j, and (v_i',v_j') will be designated by S_i, S_j, and S, respectively.

Observe that Q_i encloses the subgraph S_i, and draw non-intersecting f-contours for the branches of S_i in the manner of method 2 of the Theorem 1. Similarly, non-intersecting contours can also be drawn for the branches belonging to the subgraph S_j. Now if we consider the subgraph of T formed by the branches belonging to S, and the branches i and j, the branches i and j may be considered as two "first tip" branches of this subgraph for which the contours Q_i and Q_j have been drawn in the manner of method 1 of the Theorem1. By this very method, non-intersecting f-contours can be drawn for the branches belonging to the subgraph S. All these contours for the tree T are, thus, of non-intersecting type, which proves the theorem.

THEOREM 3. If branches i and j are selected arbitrarily in any tree T and if f-contours Q_i and Q_j are drawn such that either of these contours encloses the other, then the rest of the f-contours for T can always be drawn in such a manner that all the f-contours are of non-intersecting

PROOF: Without loss of generality, Q_j is assumed to enclose Q_i.

Non-intersecting f-contours for all the branches of T can now be obtained by the following standard procedure.

Let (v_i, v_i') and (v_j, v_j') be the vertices of the branches i and j, respectively, such that v_i and v_j are the terminal vertices of a path in T which has i and j as its terminal branches. Obviously, Q_i must enclose v_i, and Q_j must enclose v_j'. The subgraphs S_i and S_j are defined as in the last theorem and the symbol S_j' will represent the subgraph of T obtained by removing the branch j from T and taking the subgraph containing the node v_j'. Note that S_i is a subgraph of S_j'.

If S_j is imagined to be removed from T, Q_j may be considered to have been drawn as the starting f-contour in the manner of method 2 of Theorem 1. By using the same method, non- intersecting f-contours can be obtained for all the branches of S_j'. Note that Q_i has been already drawn.

Now if we consider the subgraph of T formed by the branches of S_j, and the branch j, the branch j may be looked upon as a "tip" branch of this subgraph, and f-contour Q_j may be taken as having been drawn according to method 1 of Theorem 1. This method can now be employed to draw non-intersecting f-contours for all the branches of S_j. This proves the theorem.

The above theorems lead to the conclusion that without loss of generality we can work only in terms of non-touching f-contours, and we may begin by drawing any two f-contours, as in Theorem 2 or as in Theorem 3; the rest of them can always be accommodated in a non-intersecting manner.

THEOREM 4. If non-intersecting f-contours are drawn for some tree T in a graph having $n+1$ vertices, one of the vertices will lie outside all of the f-contours and each of the remaining n vertices will correspond (Definition8) to a distinct f-contour.

PROOF: In this proof, the vertices of a branch, say j, of the tree T will be indicated by v_j and v_j'. Also, the vertex without the prime, that is v_j, will be assumed to be the one enclosed by the contour Q_j.

Consider any branch, say branch j, of the tree T and the corresponding f-contour Q_j (Fig.1). As per our notation, vertex v_j is inside the contour Q_j. In addition to j, let branches k, l, m, etc. be incident on v_j. Because all the f-contours are assumed to be of non-intersecting type, vertices v_k', v_l', v_m' etc. (that is, the vertices lying outside Q_k, Q_l, Q_m, etc., respectively) must be coincident with the vertex v_j. We can easily see that it is only the vertex v_j which must correspond to the contour Q_j. Similarly, v_k can be shown to be the vertex corresponding to Q_k, and this can be extended to the case of all the other branches of T incident on v_j.

If no branch of T other than j is incident on v_j', the vertex v_j' clearly does not lie inside any of the f-contours. However, if some branches, say p, q, r, etc., are incident on v_j'; additional f-contours corresponding to these are also to be considered. The contours Q_p, Q_q, Q_r, etc. may either be such that none of them encloses Q_j (Fig.1a), or any one of them, say Q_p, might have been drawn enclosing Q_j (Fig.1b). In the case of Fig.1a, v_j' is obviously the vertex which is not enclosed by any of the f-contours, and each of the other vertices of the graph will correspond to a distinct f-contour.

In the case of Fig.1b, discussions similar to those given above with regard to the vertex v_j' and the branches incident on it, are applicable to the vertex

v_p'. Thus,we must eventually have a single vertex which does not lie inside any of the f-contours,and each of the other vertices must correspond to a distinct f-contour. This proves the theorem.

IV. REALIZATION OF CUT-SET MATRICES

The given matrix C will be assumed to be given in the form

$$C = [U \quad C_c]$$

where U is a unit matrix of order n and C_c is the so called characteristic part of the matrix C.

A. DEFINITION AND PROPERTIES OF MATRIX W

A new matrix W is defined below,and is followed by discussions of some of its properties.

DEFINITION 10. For a given matrix C,the matrix W is defined as a square matrix of order n whith its entry W_{ij} $(i,j = 1,2,\ldots,n;\ i \neq j)$ consisting of the edges belonging to $\theta_{i,j}$ provided that the Property 1 holds for the rows i and j of the matrix C. The diagonal entries of matrix W are not defined because they would not be useful for our purpose.

If the Property 1 is not satisfied for even a single pair of rows,say rows p and q,of the matrix C,the corresponding entry W_{pq} is defined to be non-existent and in such a case the matrix C must be non-realizable.

THEOREM 5. Suppose in matrix W some rows p and q,and some column k are such that at least one edge,say r,contained in the entry W_{pk} is not contained in W_{pk}. Similarly,at least one edge,say s,of the entry W_{qk} is not contained in W_{pk}. If non-intersecting contours are drawn for the rows p,q,and k,the contour Q_k can lie only in the region between the contours Q_p and Q_q.

PROOF: The non-intersecting contours Q_p and Q_q can have only two essentially different configurations which are shown in Figs.2a and 2b (the case of Q_p enclosing Q_q is similar to the one shown in Fig.2b). The figures also show the edges r and s,both of which must also be cut by Q_k because they both appear in column k of the matrix W. Simple reasoning shows that Q_k must lie only in the region between Q_p and Q_q. This proves the theorem.

COROLLARY 5.1 Suppose the rows p,q and column k of matrix W satisfy the conditions of Theorem 5 and,in addition to that,let W_{pk} and Q_{qk} have at least a single common edge. Then Q_k,in addition to being in the region between Q_p and Q_q,would necessarily enclose either Q_p or Q_q - but not both.

COROLLARY 5.2 Suppose the rows p,q and column k of matrix W satisfy the conditions of Theorem 5 and,in addition to this,assume that W_{pq} is non-null and there is no edge common to W_{pk} and Q_{qk}. Then Q_k,in addition to lying in the region between Q_p and Q_q,would enclose either both or none of the contours Q_p and Q_q.

The above two corollaries can be easily proved with the help of Fig.2.

It will be convenient to designate a set of three contours satisfying the conditions of Corollary 5.1 by the symbol S_A,similarly a set of three contours satisfying the conditions of Corollary 5.2 will be designated by the symbol S_B.

B. REALIZATION TECHNIQUE FOR MATRIX C

The nature of the technique is such that the following terminology would prove to be very convenient.

The plot of a number of non-intersecting contours (along with the edges cut by them) pertinent to a matrix C will be called the "contour diagram" for the matrix C.

When a certain contour is to be added to the contour diagram,this particular contour will be called the "incoming" contour. An incoming contour,say Q_p,will always be drawn such that those edges of θ_p which already exist in the contour diagram are cut by it,and only the remaining edges in θ_p would be additionally drawn cutting the contour Q_p. It will be assigned such an orientation that its orientatiin with respect to a directed edge k,already drawn in the contour diagram,becomes similar (opposite) if the entry in the p-th row and k-th column of the matrix C is 1 (-1). Next a test called "sign test for the contour Q_p" would be applied to verify the necessary condition that the mutual orientation of Q_p and each of the other contours already existing in the contour diagram must be as specified in the matrix C. In case the sign test fails,the matrix C must be non-realizable.

If the contours existing in the contour diagram are such that one of them, say Q_q,encloses all the rest,and if an incoming contour Q_p is desired to be drawn in the region R_q' then,without loss of generality,Q_p would always be drawn enclosing the contour Q_q. A contour Q_p drawn in this manner will be indicated by the symbol $Q_p(R_q')$.

The symbol for a contour will carry a superscript which would indicate its number in the serial order in which the various contours are drawn in the contour diagram. Thus,

$$Q_p^1 \, , \, Q_r^2 \, , \, Q_s^3 \, ,\text{etc.}$$

would indicate that the first,second,third,etc. contours to be drawn in the contour diagram are Q_p,Q_r,Q_s,etc.,respectively.

As a matter of further notation,let Q_k be one of the contours already existing in the contour diagram and let Q_p be the incoming contour. If Q_p is drawn in the region r_k but without itself enclosing any of the already existing contours,it will be indicated by the symbol $Q_p(r_k)$. It may,however, be noted that some of the already existing contours may exist in the region between Q_k and $Q_p(r_k)$. On the other hand,if Q_p is drawn in the region r_k in such a manner that,after it has been drawn,each of the contours $Q_r,Q_s,\ldots,$ Q_t (all of them already existing in the contour diagram) would appear to exist in the region r_p,then,contour Q_p drawn in this manner would be indicated by the symbol $Q_p(r_k;\text{Encl.}Q_r,Q_s,\ldots,Q_t)$. It may be noted that contours other than $Q_r,Q_s,\ldots,Q_t$ may happen to be existing in the region R_p,but in the parenthesis we do not show those contours which may be present in the region $R_r,R_s,\ldots,R_t$.

If the contour Q_p is to be added in some step j,the symbol S_j will be used to indicate the set of all the different topological ways in which Q_p could be drawn satisfying the conditions regarding the edges cut and the sign test.

The proposed realization technique can now be described as follows:

(i) Write down the matrix W for the given matrix C. If matrix W is found to be non-existent,the matrix C must be non-realizable.

(ii) Matrix W is scanned to find if there exists a set S_A or S_B. If such a set has been found out,the contour diagram is started by drawing the three contours represented by it,otherwise the contour diagram is started by drawing any two contours having at least one common edge. Other things being equal,precedence for being drawn in the contour diagram is given to a contour having a comparatively large number of edges common with the other contours. This imparts a somewhat rigid frame of structure to the contour diagram which would admit a comparatively smaller number of different possibilities in which a subsequent contour might be assigned acceptable positions in the contour diagram.

(iii) Another contour is found - either by inspecting the remaining rows in matrix C or by using matrix W - which has at least one edge common with the ones already drawn,and which could be assigned the minimum number of different acceptable positions (ideally only a single one) in the contour diagram. This contour is now drawn in any one of these acceptable ways. It may be stressed that the topological requirements are usually so stringent that in all but the simplest cases,the possible number of different topological layouts admissible for an incoming contour is either restricted to one,or a very small number indeed.

If a contour having at least one edge common with the ones already drawn cannot be found,the matrix C - if realizable - must correspond to a "hinged" graph which can also be obtained by the present technique.

(iv) The above process is repeated till the complete contour diagram has actually obtained;or proved to be impossible (after considering all the relevant different possibilities) which proves that matrix C is non-realizable.

(v) If the complete contour diagram has been obtained above,the desired graph can be easily obtained either by coalescing together the appropriate vertices of the edges in the contour diagram itself,or by first obtaining the pertinent node-branch incidence matrix from the plot.

C. ILLUSTRATIVE EXAMPLE

Determine the realizability of the matrix C given in Fig.3 (zeros have been omitted for convenience) as a f-cut-set matrix.

STEP 1. The matrix W corresponding to the given matrix C can be written down as shown in Fig.4.

STEP 2. We now wish to find if there exists any set S_A or S_B in the matrix W. The first column of W does not indicate any set S_A or S_B;the second column,however,shows that Q_1,Q_5,and Q_2 form a set S_A. Hence,the contour diagram (Fig.5) is started by drawing the contours

$$Q_1^1 \text{ , } Q_5^2(R_1') \text{ , and } Q_2^3(r_5;\text{Encl. } Q_1).$$

The contour Q_1 is arbitrarily oriented as shown in the figure,and suitable orientations for other contours,as subsequently obtained,are also shown there.

STEP 3. It is easily seen by scanning the matrix W that no further set S_A or S_B is formed by any two of the already drawn contours in conjunction with a third one not drawn so far. However,by considering the remaining rows in the matrix C and the contour diagram,it is easily found that the contour diagram would admit only a single topological configuration for the contour Q_3. Hence we have

$$S = Q_3^4(r_2).$$

STEP 4. No special preference for being selected next can be given to either of the remaining two rows. However,the 4-th is selected now as Q_4 has two edges common with other contours while the other,namely Q_6,has only one. It is found that the contour Q_4 can be drawn in two different ways. These possibilities are written below as the elements of the set S_4.

$$S = \left\{ \underset{\text{(step 5)}}{Q_4^5\,(r_1)} \quad , \quad Q_4^5\,(R_5') \right\}$$

The "(step 5)" written above indicates that contour $Q_4^5(r_1)$ is the one actually drawn in the present step and the contour diagram,so obtained,is considered further in Step 5.

STEP 5. The remaining contour Q_6 is added in this step giving

$$S = \left\{ \underset{\text{(selected)}}{Q_6^6\,(r_4)} \quad , \quad Q_6^6\,(R_5') \right\}$$

It follows that the given matrix C is indeed realizable as a f-cut-set matrix.

STEP 6. The graph realizing the given matrix C as a f-cut-set matrix can now be easily obtained from the complete contour diagram. It is shown in Fig.6.

V. CONCLUSIONS

A new technique for realization of a given matrix as a cut-set matrix of some appropriate graph (directed or non-directed) has been presented in this paper. Necessary and sufficient conditions for such a realization are implied in the realization technique.

The method given here is entirely due to the systematic applications of properties of f-contours developed in the paper. The unique approach has resulted in a simple,systematic,and yet perfectly general realization technique. The procedure is based directly on topological considerations, and thus keeping in view the topological nature of the problem itself,it is not surprising that the technique has come out to be an extremely effective one. The new method involves no calculations and the given matrix is not required to be subjected to any additional manipulations either prior to or during the realization process.

ACKNOWLEDGEMENT

The authors are very grateful to Dr.A.K.Kamal,Head of the Department of Electronics and Communication Engineering,University of Roorkee,for his kind interest and encouragement.

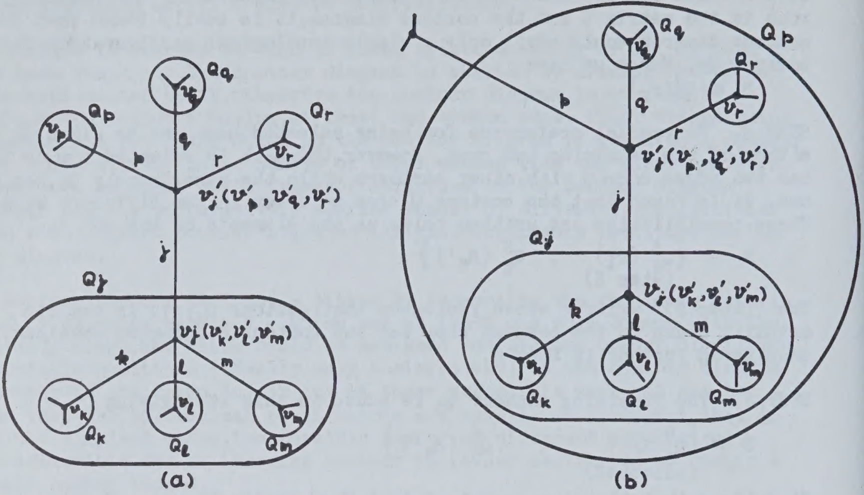

Q_q
v_q
Q_p
v_p
Q_r
v_r
p
q
r
$v_j'(v_p', v_q', v_r')$
j
Q_j
$v_j(v_k', v_l', v_m')$
k
l
m
v_k
v_l
v_m
Q_k
Q_l
Q_m
(a)
(b)

FIG. 1 - SET OF f - CONTOURS

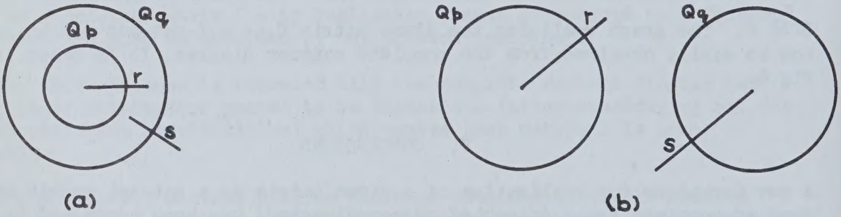

Qq
Qp
r
s
(a)
Qp
r
Qq
S
(b)

FIG. 2- CONFIGURATIONS WHICH TWO NON-INTRESECTING CONTOURS CAN ASSUME.

$$C = \begin{array}{c|ccccccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 1 & 1 & & & & & & 1 & 1 & & 1 & \\ 2 & & 1 & & & & & -1 & -1 & -1 & -1 & 1 \\ 3 & & & 1 & & & & & & 1 & & -1 \\ 4 & & & & 1 & & & 1 & 1 & & & \\ 5 & & & & & 1 & & -1 & -1 & -1 & & \\ 6 & & & & & & 1 & -1 & & & & \end{array}$$

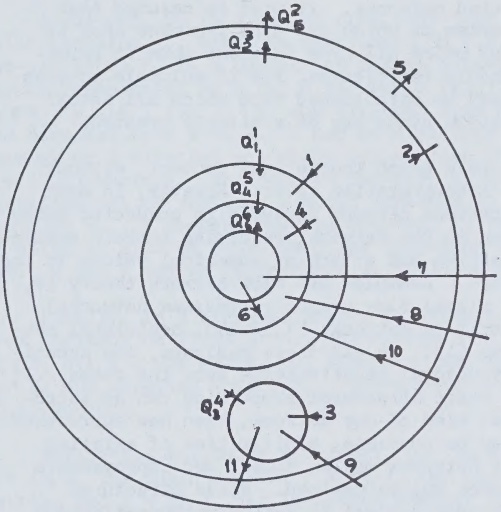
FIG. 3

f CUT-SET →

f-CUT-SET ↓

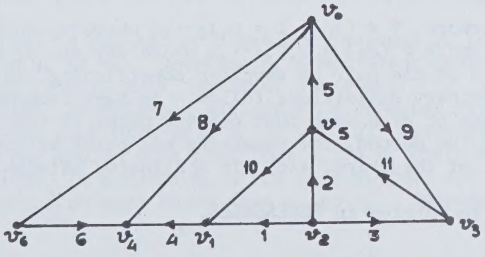
$$W = \begin{array}{c|cccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & & \substack{7\\8\\10} & 0 & \substack{7\\8} & \substack{7\\8} & 7 \\
2 & \substack{7\\8\\10} & & \substack{9\\11} & \substack{7\\8} & \substack{7\\8\\9} & 7 \\
3 & 0 & \substack{9\\11} & & 0 & 9 & 0 \\
4 & \substack{7\\8} & \substack{7\\8} & 0 & & \substack{7\\8} & 7 \\
5 & \substack{7\\8} & \substack{7\\8\\9} & 9 & \substack{7\\8} & & 7 \\
6 & 7 & 7 & 0 & 7 & 7 &
\end{array}$$

FIG. 4



Q_5^2
Q_2^3
5
2
Q_1^1
1
Q_4^5
4
Q_6^6
7
8
6
10
Q_3^4
3
11
9

FIG. 5
THE COMPLETE CONTOUR DIAGRAM FOR EXAMPLE-1



v_0
7
8
5
v_5
9
10
2
11
v_6
6
v_4
4
v_1
1
v_2
3
v_3

FIG. 6
THE GRAPH REALIZING THE MATRIX C GIVEN IN EXAMPLE - 1

SOME DUALITY RELATIONSHIPS IN NETWORKS

M. I. DESSOUKY
Associate Professor of Industrial Engineering
Department of Mechanical and Industrial Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

E. J. DUNNE
Captain, United States Air Force
Wright-Patterson Air Force Base
Dayton, Ohio 45433

INTRODUCTION

Many engineering and management systems have been represented by network based models for the purpose of their analysis and design. A network is a connected graph with no loops (a loop is an edge or arc with the same vertex at its two ends). It may be directed or undirected, although the emphasis in this work is on directed networks. *It will be assumed that each network has one source (a vertex at which all arcs incident upon it emanate) and one sink (a vertex at which all arcs incident upon it terminate).* This assumption is not unduly restrictive, for if multiple sources exist, a single "master source" may be established from which all actual sources immediately follow. Multiple sinks may be similarly treated.

A network is thus definable as a graph theoretical concept, without necessarily attaching a physical interpretation to it. However, in most engineering and management applications network analysis is conducted after a specific interpretation is given to the network, requiring certain assumptions about some additional properties and attaching numerical values to the arcs and/or vertices of the network. Examples are flow network theory [8, 9]*, distance networks [2,6,11], signal flow graphs (sometimes networks) [6], electrical networks [10], activity networks [7,12,14], decisional networks [3], and simulation networks [1]. In all these analyses, the numerical rather than the logical or structural relationships were the focus. It is reasonable to assume that, if basic structural properties can be established between the fundamental entities of any network, then new approaches to existing network techniques may be possible, similarities of existing techniques for different types of networks may be noted, and improvements in currently used network approaches may be devised. While structural properties of general graphs have been subject to extensive investigation [2,13], very little attention was given to those unique to networks.

The exploration of some of the structural properties of networks, particularly certain duality relationships, is the objective of this paper.

BASIC NETWORK CONCEPTS

Let N(V,A) define a network, $V = (V_i)$, $i = 0,1,\cdots,n$ where v_o is the source and v_n is the sink, and $A = (a_\ell) = 1,2,\cdots,m$ where any arc a_ℓ may also be referred to in terms of the pair of vertices constituting its end vertices (v_i, v_j). Some pertinent definitions follow. An *Edge Progression* $G = (G_j)$ is a finite sequence of arcs such that one end point of the first arc is also an end point of the second, the remaining end point of the second is also an end point of the third, etc. In a directed network G may be represented by a (-1,0,1)-vector (g_ℓ) such that $g_\ell = 0$ for $a_\ell \notin G$, $g_\ell = 1$ for $a_\ell \in G$ and directed forward along G, and $g_\ell = -1$ for $a_\ell \in G$ and directed backward or reverse. In undirected networks and in directed ones, where arc orientation is not of concern, G is represented as a (0,1)-vector. If the first and last vertices of the sequence are identical, it is called a *circuit* $R_d = (r_{d\ell})$ such that $r_{d\ell} = 0$ for $a_\ell \notin R_d$, $r_{d\ell} = 0$ for $a_\ell \in R_d$ and running counterclockwise with respect to R_d, and $r_{d\ell} = -1$ for $a_\ell \in R_d$ and running counterclockwise.

---

*Numbers in brackets refer to entries in REFERENCES.

If the first and last vertices of the sequence are distinct, and there are no repeated arcs or vertices, the edge progression is called a chain. Of particular interest are chains and edge progressions from v_o to v_n, defined by (-1,0,1)-vectors or (0,1)-vectors H_i, G_j, respectively.

In a directed network, an *arc progression* is an edge progression where all arcs have the same orientation as the progression. It is always expressed as a (0,1)-vector. Similarly defined are cycles, which are a special case of circuits, and paths, which are a special case of chains. Let P_i denote a path from v_o to v_n.

A *cut-set*, C, is a set of arcs that join a specified set of vertices W which contains v_o, with its complement in V,$\overline{W}$, which contains v_n. A *proper cut-set* (PCS) K_j is a cut-set no proper subset of which is a cut-set. K_j may be represented as a (-1,0,1)-vector ($k_{j\ell}$) such that $k_{j\ell} = 0$ for $a_\ell \notin K_j$, $k_{j\ell} = 1$ for $a_\ell \in K$ and directed from W to $\overline{W}$, and $k_{j\ell} = -1$ for $a_\ell \in K_j$ and directed from $\overline{W}$ to W. C may be similarly represented. In an undirected network or a directed one where orientation of arcs is neglected, C and K are (0,1)-vectors. An *oriented cut-set*, C, is a set of arcs directed from a specified set of vertices W which contains v_o to its complement in V,$\overline{W}$, which contains v_n. A *proper oriented cut-set*, K, (POCS) is an OCS no proper subset of which is also an OCS. Ford and Fulkerson's [8] "cut" is an oriented cut-set. $\vec{C}$ and $\vec{K}$ may be represented as (0,1)-vectors.

Finally, in a directed graph, a vertex $v_i \in V$ may be represented by the arcs incident on it as a (-1,0,1)-vector $B = (b_{i\ell})$, such that $b_{i\ell} = 0$ for $a_\ell \notin B$, $b_{i\ell} = 1$ for $a_\ell \in B_i$ and directed away from v_i, and $b_{i\ell} = -1$ for $a_\ell \in B_i$ and directed toward v_i. Where orientation is not relevant, B_i is a (0,1)-vector.

Define the following matrices: $\underline{\underline{G}} = \{G_i\}$, the edge progression matrix; $\underline{\underline{H}} = \{H_i\}$, the chain matrix; $\underline{\underline{P}} = \{P_i\}$, the path matrix; $\underline{\underline{R}} = \{R_d\}$, the circuit matrix; and $\underline{\underline{B}} = \{B_i\}$, the incidence matrix. Similarly, let $\underline{\underline{C}} = \{C\}$, $\underline{\underline{\vec{C}}} = \{\vec{C}\}$, $\underline{\underline{K}} = \{K\}$ and $\underline{\underline{\vec{K}}} = \{\vec{K}\}$.

BASIC NETWORK PROPERTIES

The following is a set of properties that relate chains, paths, cut-sets, OCS's, PCS's and POCS's to each other. They are represented as theorems without proof for lack of space, but proof may be found where references are given.

Theorem (1a) *A cut-set of a network includes at least one arc on each chain from* v_o *to* v_n. *Conversely, a set which includes at least one arc on each chain from* v_o *to* v_n *contains a cut-set.* [5]

Theorem (1b) *An oriented cut-set of a directed network includes at least one arc on each path from* v_o *to* v_n. [8] *Conversely, a set which includes at least one arc on each path from* v_o *to* v_n *contains an oriented cut-set.* [5]

Theorem (2a) *In a network, a* PCS *is a minimal set of arcs which includes at least one arc on each chain from* v_o *to* v_n. *Conversely, a minimal set of* *arcs which includes at least one arc on each chain from* v_o *to* v_n *is a* PCS [5].

Theorem (2b) *In a directed network, a* POCS *is a minimal set of arcs which includes at least one arc on each path from* v_o *to* v_n*, minimal in the sense that the deletion of any arc from the set leaves a set which includes no arcs from at least one path from* v_o *to* v_n*. Conversely, a minimal set of arcs which includes at least one arc on each path from* v_o *to* v_n *is a* POCS. [5]

Theorem (3a) *In a network, given a cut-set* $(W,\overline{W})$*, the necessary and sufficient conditions for* $(W,\overline{W})$ *to be a* PCS *are that all* $w \in W$ *be connected to* v_o *by members of* W *and all* $\overline{w} \in \overline{W}$ *be connected to* v_n *by members of* $\overline{W}$ [5]. *Stated differently,* $(W,\overline{W})$ *is a* PCS *if and only if both* W *and* $\overline{W}$ *are connected sets after the removal of* $(W,\overline{W})$.[2]

Theorem (3b) *In a directed network, given* S *and* T *as the set of starting and terminal vertices of the arcs of the oriented cut-set* $(W \rightarrow \overline{W})$*, respectively, (i.e.,* $S \in W$ *and* $T \in \overline{W}$*), the necessary and sufficient conditions for* $(W \rightarrow \overline{W})$*, to be a* POCS *are that all* $s \in S$ *be path-connected to* v_o *by members of* W *and* v_n *be path-connected to all* $t \in T$ *by members of* $\overline{W}$ [5]. (A vertex or an arc g is path-connected to a vertex or an arc f if a path exists from f to g.)

Theorem (4) *In a directed network, if* $(W \rightarrow \overline{W})$ *is a* POCS, *then* $(W,\overline{W})$ *is a* PCS. [5]

This section has developed some basic network properties of POCS's and PCS's. **Theorems (1a)** and **(2a)** establish the basic relation between a PCS and the chains of a network. **Theorems (1b)** and **(2b)** establish similar relations between a POCS and the paths of a network. **Theorems (3a)** and **(3b)** develop identifying properties for PCS's and POCS's, respectively, and **Theorem (4)** demonstrates the relationship between POCS's and PCS's.

DUALITIES IN (0,1)-MATRICES

Special insights may be gained when networks are represented as (0,1)- or (-1,0,-1)-matrices. In this section an important property of (0,1)-matrices is discussed and interpreted as it applies to networks.

In a (0,1)-matrix $\underline{\underline{A}}$, row A_i is *dominating* if there exists another row A_j such that $A_i \geq A_j$. A *proper matrix* is a (0,1)-matrix with no dominating rows. The following matrices obviously are proper: $\underline{\underline{H}}$, $\underline{\underline{P}}$, $\underline{\underline{K}}$, $\underline{\underline{\vec{H}}}$, and $\underline{\underline{\vec{K}}}$.

A cover C of an (nxm) (0,1)-matrix $\underline{\underline{A}}$, $C(\underline{\underline{A}})$, is an m-dimensional (0,1)-vector such that $A_i \cdot C \geq 1$, all i. A *proper cover* Q of a (0,1)-matrix $\underline{\underline{A}}$, $Q(\underline{\underline{A}})$, is a cover such that no other cover $C(\underline{\underline{A}})$ exists and $Q \geq C$.

A significant property of proper (0,1)-matrices is the following: [5] For any proper (0,1)-matrix $\underline{\underline{A}}$ and the matrix of all proper covers of $\underline{\underline{A}}$ (put in matrix form), $\underline{\underline{Q}}(\underline{\underline{A}})$, $\underline{\underline{Q}}[\underline{\underline{Q}}(\underline{\underline{A}})] = \underline{\underline{A}}$; *i.e.,* matrices $\underline{\underline{A}}$ and $\underline{\underline{Q}}(\underline{\underline{A}})$ are duals in the sense that each is the set of all proper covers of the other.

Taking **Theorems (2a)** and **(2b)** and considering jointly the dualizations just mentioned, these network relations may be concisely stated as:

In a network, a set of arcs is a

$$\left\{\begin{matrix}\text{POCS}\\ \text{PATH}\\ \text{PCS}\\ \text{CHAIN}\end{matrix}\right\} \begin{matrix}\text{if and only if it is a minimal set of arcs}\\ \text{containing at least one arc from every}\end{matrix} \left\{\begin{matrix}\text{PATH}\\ \text{POCS}\\ \text{CHAIN}\\ \text{PCS}\end{matrix}\right\}$$

of the network.

The duality between chains and PCS's has been observed in planar graphs and was represented by the dual graph, but it has not been established in the general sense reported here.

NETWORK DUALITIES IN (-1,0,1)-MATRICES

In certain types of network problems, the designation of the orientation (sign) of the individual arcs in some of the sets (vectors) is essential to the solution of the problem. Examples are maximum flow and minimum-cost flow problems in directed networks, where flow-augmenting paths may be thought of as flow-carrying chains. Another example is cost-duration analysis in activity networks where, in one approach [4,5], step-by-step minimum-cost reduction of the project duration requires the choice, at every step, of a minimum-cost PCS, whose oriented arcs represent activities which may be decreased in duration.

In these cases, the use of (-1,0,1)-vectors is required. The following theorems establish some duality properties expressed in (-1,0,1)-matrix formulation.

Theorem (5) *In a directed network, given* $\underline{\underline{H}}$ *and* $\underline{\underline{K}}$, $\underline{\underline{H}} \cdot \underline{\underline{K}}^T = \underline{\underline{U}}$, *where* $\underline{\underline{U}} = \{u_{ij}\}$, $u_{ij} = 1$, *all* i,j.

Proof Given any chain H_i and any PCS $K_j = (W,\overline{W})$, as H_i is traced from v_o to v_n, it will cross from W (including v_o) to $\overline{W}$ (including v_n) at least once. The arc a_f on H_i making this crossing may be oriented (forward) with respect to H_i, in which case $h_{if} = 1$, or unoriented (reverse) with respect to H_i, in which case $h_{if} = -1$. In the former case, a_f will be an oriented member of K_j with $k_{jf} = 1$, since a_f is directed from W to $\overline{W}$ and, in the latter case, a_f will be an unoriented member of K_j with $k_{jf} = -1$, since a_f is directed from $\overline{W}$ to W. In either case, $h_{if} \cdot k_{jf} = 1$, any f.

If H_i makes the crossing from W to $\overline{W}$ only once, then the only common arc between H_i and K_j is a_f. This means that there is only one arc a_f for which both vectors H_i and K_j have nonzero entities. Hence, $H_i \cdot K_j^T = h_{if} \cdot k_{jf} = 1$.

Suppose H_i crosses from W to $\overline{W}$ more than once, say F times, on a set of arcs $\{a_f\}$. Before every new crossing from W to $\overline{W}$ beyond the initial one, H_i must return from $\overline{W}$ to W on some arc. Let the number of crossings from $\overline{W}$ to W which H_i makes, on some set of arcs $\{a_g\}$, be G. Since H_i starts at v_o in W and ends at v_n in $\overline{W}$, $F - G = 1$.

Any arc a_g on H_i making a crossing from $\overline{W}$ to W may be oriented (forward) with respect to H_i, $h_{ig} = 1$, or unoriented (reverse) with respect to H_i, $h_{ig} = -1$. In the former case, a_g will be an unoriented member of K_j, $k_{jg} = -1$, since a_g is directed from $\overline{W}$ to W, and, in the latter case, a_g will be an oriented member of K_j, $k_{jg} = 1$, since a_g is directed from W to $\overline{W}$. In either case, $h_{ig} \cdot k_{jg} = -1$, any g.

The only arcs which have nonzero entries in both vectors H_i and K_j are those which are members of either $\{a_f\}$ or $\{a_g\}$. Therefore,

$$H_i \cdot K_j^T = \sum_{\text{all } f} h_{if} \cdot k_{jf} + \sum_{\text{all } g} h_{ig} \cdot k_{jg}, = F - G = 1, \text{ all } i,j.$$

Therefore, $\underline{\underline{H}} \cdot \underline{\underline{K}}^T = \underline{\underline{U}}$, where $\underline{\underline{U}} = \{u_{ij}\}$, $u_{ij} = 1$, all i,j, and the theorem is proved.

Some interesting results follow from this theorem. First, define a *basic cut-set* (BCS) as a cut-set containing no unoriented arcs, that is, arcs directed from $\overline{W}$ to W.

Corollary 1 *In a directed network, given any path and any* BCS, *the two will have one and only one arc in common.*

Corollary 2 *In a directed network,* $\underline{\underline{G}} \cdot \underline{\underline{C}}^T = \underline{\underline{U}}$, *where* $\underline{\underline{U}}$ *is defined as above.* This corollary is a generalization of the theorem and can be proved in the same way.

Following are some relationships between the circuit matrix and each of the proper cut-set and the incidence matrices.

Theorem (6) $\underline{\underline{R}} \cdot \underline{\underline{K}}^T = \underline{\underline{0}}$, *where* $\underline{\underline{0}}$ *is a null matrix.* The proof of this corollary is derived from the fact that, for any R_d and any K_j, R_d must either pass on vertices exclusively in W or in $\overline{W}$ and, in either case, it will not intersect K_j and $R_d \cdot K_j^T = 0$ or it will cross from one set (W or $\overline{W}$) to the other. In the latter case, the number of crossings from W to $\overline{W}$,F, must be equal to the number of crossings from $\overline{W}$ to W,G, and $R_d \cdot K_j^T = 0$ again.

Theorem (7) $\underline{\underline{R}} \cdot \underline{\underline{B}}^T = \underline{\underline{0}}$.

Proof For any R_d and B_i, if $v_i \notin R_r$, then $R_d \cdot B_i^T = 0$ and, if $v_i \in R_d$, then B_i shares two arcs with R_d such that the sign of the product $(r_{d\ell} \cdot b_{\ell i})$ for the first arc is different from that of the product for the second arc and $R_d \cdot B_i^T = 0$ also.

Parallel results for some of the above theorems and corollaries may be derived for (0,1)-vectors and matrices with the use of Boolean or Mod 2 operations for addition which specify: $0 + 0 = 0$, $0 + 1 = 1 + 0 = 1$, $1 + 1 = 0$ (Mod 2). For any two (-1,0,1)-vectors X and Y, let the operation $(X^a * Y^{aT})$ mean the vector multiplication of X^a and Y^a, (0,1)-vectors whose elements are the absolute values of X and Y, respectively, with Boolean addition of the products of corresponding elements. It is obvious that $X^a * Y^{aT} = (X \cdot Y^T)^a$.

Proofs of the (0,1) version of Theorems (6) and (7) are shown in the literature [2,13]. A summary of the relationships in Theorems (5), (6) and (7) is given in Fig. 2.

The following theorem establishes the uniqueness of the relationship in Theorem (5).

Theorem (8) *In a directed network,*

(a) *Given* $\underline{H}$ *and* X, a(-1,0,1) *m-dimensional vector,* $X \in \underline{\underline{K}}$ *if and only if* $H_i \cdot X^T = 1$, *all* i, *and no* $X' \subset X$ *for which* $H_i \cdot X' = 1$, *all i, and*

(b) *Given* $\underline{K}$ *and* Y, a(-1,0,1) *m-dimensional vector,* Y $\underline{\underline{H}}$ *if and only if* $Y \cdot K_j^T = 1$, *all* j, *and no* $Y' \subset Y$ *for which* $Y' \cdot K_j^T = 1$, *all* j.

NOTE: (For (-1,0,1)-vectors, $X' \subset X$ means that all zero elements of X are zero elements of X', at least one nonzero element of X is a zero element of X', and all nonzero elements of X' are identical to the corresponding elements of X. In other words, $X' \subset X \rightarrow X'^a \subset X^a$.)

This theorem states the necessary and sufficient conditions for membership in $\underline{\underline{K}}$ and $\underline{\underline{H}}$, in terms of (-1,0,1)-vectors and matrices.

Proof For any m-dimensional (-1,0,1)-vector X, it follows from the previous section that, if $X \in \underline{\underline{K}}$, then

(a) $H_i^a \cdot X^{aT} \geq 1$, all $H_i \in \underline{\underline{H}}$.

(b) There is no $X' \subset X$ for which condition (a) applies.

For any $X \in \underline{\underline{K}}$, Theorem (5) asserts that $H_i \cdot X^T = 1$, all H_i. No $X' \subset X$ exists for which $H_i \cdot X'^T = 1$, all H_i; otherwise, $H_i^a \cdot X'^{aT} \geq 1$, all H, and condition (2) is violated for X. *NOTE:* (This statement is true since, for any two scalar values x and y, $x^a \cdot y^a \geq x \cdot y$, where x^a and y^a represent the absolute values of x and y, respectively.)

Now let X be a (-1,0,1) m-dimensional vector satisfying (a') $H_i \cdot X^T = 1$, all $H_i \in \underline{\underline{H}}$, and (b') no $X' \in X$ exists for which (a') is true.

X^a is a cover of $\underline{\underline{H}}^a$; hence, it is the absolute value of a cut-set and either $X \in \underline{\underline{C}}$, the set of cut-sets or $\exists Y \in \underline{\underline{C}}$ ($Y^a = X^a$ and $Y \neq X$). From Corollary (2), Theorem (5), statement (a') applies to Y as it does to X.

Let $X' = (X + Y)/2$. From (A') and the similar relation for Y, $H_i \cdot X'^T = 1$, all $H_i \in \underline{\underline{H}}$. Also, from $Y^a = X^a$, X' must be a (-1,0,1)-vector, with dimension m, such that $X' \subset X$, violating (b'). This is true because, if an element in X is identical to its counterpart in Y, then the corresponding element in X' is identical to both; if they are different in sign, then the corresponding element in X' is zero. Therefore, $X \in \underline{\underline{C}}$.

Either $X \in \underline{\underline{K}}$ or some other vector $Z \subset X$ and $Z \in \underline{\underline{K}}$. In the latter case, from Theorem (5), (a') applies to Z, contradicting (b') as it applies to X. Hence, $X \in \underline{\underline{K}}$. This completes the proof of statement (1) of the theorem.

SUMMARY

In this paper, the most significant developments in identifying network structural relationships are:

(1) A complete dualization exists between a network's paths and its POCS's.
(2) A complete dualization exists between a network's chains and its PCS's.
(3) The scalar product of any (-1,0,1)-vector representing a network chain with a vector representing a PCS will always equal unity.
(4) Some important relationships exist in any network among the following (-1,0,1)-matrices: $\underline{\underline{H}}$, the chain matrix; $\underline{\underline{K}}$, the PCS matrix; G, the edge progression matrix; $\underline{\underline{C}}$, the cut-set matrix; $\underline{\underline{R}}$, the circuit matrix; and $\underline{\underline{B}}$, the incidence matrix.

The application of these new concepts may be limited only by the different management, engineering, and other systems which can be modeled with networks. For example, the problems of maximum flow in a directed network is also the problem of finding the minimum-valued POCS. Two applications already developed [4,5] present new approaches to the solution of the cost-duration analysis problem.

REFERENCES

1. J. M. Antill and R. W. Woodhead, Critical Path Methods in Construction Practice, John Wiley and Sons, Inc., 1965.

2. R. G. Busacker and T. L. Saaty, Finite Graphs and Networks: An Introduction with Applications, McGraw-Hill Book Co., 1965.

3. W. Crowston and G. L. Thompson, "Decision CPM: A Method for Simultaneous Planning, Scheduling, and Control of Projects," Operations Research, 15 (1967), No. 3, pp. 407-426.

4. M. I. Dessouky and E. J. Dunne, "Cost-Duration Analysis with the Cut Network," AIIE Trans., III (June 1971), No. 2, pp. 46-55.

5. E. J. Dunne, "Network Theory and Project Resource Management," Ph.D. Thesis, Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, 1971.

6. S. E. Elmaghraby, Some Network Models in Management Science, Lecture Notes in Operations Research and Mathematical Systems, Springer-Verlag, 1970.

7. J. W. Fondahl, "A Non-Computer Approach to the CPM for the Construction Industry," Department of Civil Engineering, Stanford University, 1962.

8. L. R. Ford, Jr., and D. R. Fulkerson, Flows in Networks, Princeton University Press, 1962.

9. T. C. Hu, Integer Programming and Network Flows, Addison-Wesley Publishing Co., 1969.

10. W. H. Ingram and C. M. Cramlet, "On the Foundations of Electrical Network Theory," J. Mathematical Physics, 23 (1944), pp. 134-155.

11. A Kaufmann, Graphs, Dynamic Programming and Finite Games, Academic Press, 1967.

12. J. R. Kelly, Jr., "Critical Path Planning and Scheduling: Mathematical Basis," Operations Research, 9 (1961), No. 3, pp. 296-320.

13. S. Seshu and M. B. Reed, Linear Graphs and Electrical Networks, Addison-Wesley, 1961.

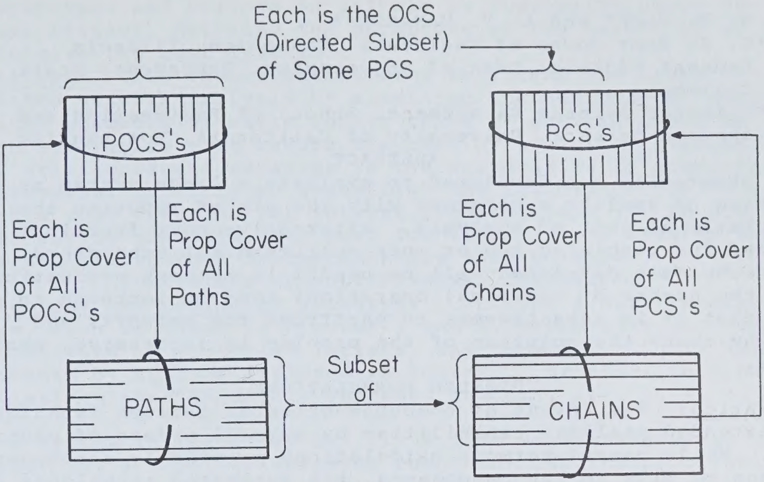

Each is the OCS
(Directed Subset)
of Some PCS
POCS's
PCS's
Each is Prop Cover of All POCS's
Each is Prop Cover of All Paths
Each is Prop Cover of All Chains
Each is Prop Cover of All PCS's
PATHS
Subset of
CHAINS

Figure 1 Duality Relationships in (0,1)-Matrices

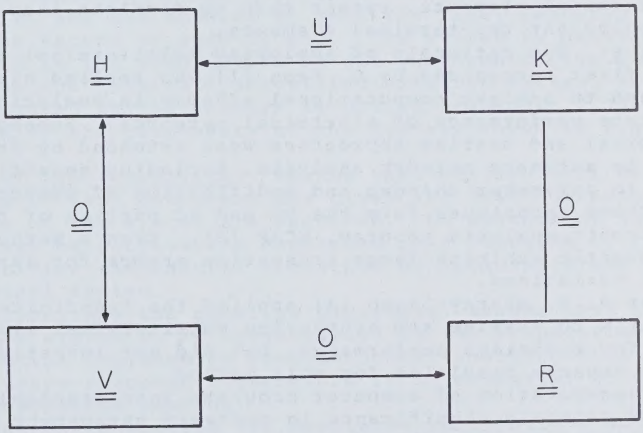


Figure 2 Duality Relationships in (-1,0,1)-Matrices

STRATEGIES FOR TEARING MULTITERMINAL NETWORKS

W. W. HAPP* and L. P. McNAMEE**
*U. S. Army Corps of Engineers, Champaign, Illinois
Present Address: Dean of Engineering, Sacramento State College
**Computer Science Department, School of Engineering and Applied Science, University of California, Los Angeles

ABSTRACT

Strategies are developed to evaluate a large system as a function of smaller subsystems with the aim of reducing the computational cost of analysis. Alternative cost functions are defined for combining two or more multiterminal subsystems. The techniques developed will be useful in network evaluation when the number of numerical operations tend to increase so rapidly that it is advantageous to partition the network, and thereby reach the solution of the problem in successive, smaller steps.

PROBLEM FORMULATIONS

Motivation: The advent of computer-oriented network techniques has extended analysis capabilities by several orders of magnitude. While manual network calculations previously encompassed a range of five to ten components, the automated techniques permitted calculations in the range of twenty to fifty components. With increased computational facilities the demand rose rapidly for capabilities: (1) To extend available techniques to networks with hundreds of components, (2) to replace conventional techniques by more versatile methods, permitting analysis of frequencies dependence components, piece-wise linear circuit elements, and eigenvalue formulations, (3) to devise descriptions of multiterminal elements, rather than to simulate them by a set of equivalent two-terminal elements.

Perspective: The rationale of employing multiterminal components was first recognized by G. Kron [1] who applied his 'tearing' method to achieve computational economy in analyzing steady state performance of electrical networks. Subsequently multiterminal and tearing approaches were extended by Branin[2] in order to automate network analysis, including sensitivity analysis to parameter changes and modification of network topology. These techniques form the DC and AC portion of the electronic circuit analysis program, ECAP [3]. Kron's method however frequently exhibits large truncation errors for large parameter variations.

Later M. A. Murray-Lasso [4] applied the indefinite addmittance matrix to examine and synthesize multiterminal electronic networks for transient performance, but did not investigate partitioning aspects resulting for this method.

The segmentation of computer programs into tractable modules is of economic significance in software management which can be extended to network reduction schemes. The problem of finding the minimum cost for a segmented program has recently been examined in terms of procedures for partitioning graphs [5-8].

Kron's principal objective in tearing a large network into smaller multiterminal components was the prospect of reducing the cost of analysis. Although a cost function in relation to the number of elements was alluded to by Kron [1], a reasonable estimate in terms of cost reduction strategies needs to be formulated. Such a strategy would aim to 'tear' the network into subsystems and perform an anlysis in successive steps on smaller systems, defining the magnitude of the partition so as to minimize the cost function.

Objectives: This investigation aims to extend the scope of automated network analysis by examining the behavior of multiterminal elements as basic constituents of networks, and to establish strategies that reduce the cost of analysis. The strategies utilize recent advances in the systematics of combining multiterminal networks to accomplish the following: (1) To establish a cost function for network combinations; (2) to propose a strategy to minimize this cost function; (3) to examine the scope and limitations in implementing this strategy.

SUBSYSTEMS OF MULTITERMINAL COMPONENTS

Terminology and Coding of Networks: Consider a multiterminal system, such as shown in Figures 1 and 2, where each component is identified by encircled letters, and common terminals between components by numbers enclosed in squares. External terminals are distinguished by doubly enclosed numbers, i.e. 1, 2, 3, and 4.

A system is specified by: (1) A list of multiterminal components, (2) a topology defining component interconnections, (3) a convention defining terminals of the system and of its components. The terminals of components are specified by: (1) a numeral which refers to the common interconnection of the terminal to other component terminals, and (2) by the position of that numeral which refers to the order of the terminal within the component based upon a specified starting point (a dot in Figures 1 and 2) and moving clockwise. For example A (1368) means that component A has four terminals connected to nodes 1, 3, 6, and 8, and are ordered so that the first terminal goes to node 1, the second to node 3, etc. The system is denoted by Z(1234) and is the combination of all components; the numerals of Z define the terminals associated with the system, and are not eliminated in the network reduction procedure. Tables 3 and 6 specify the components of the systems shown in Figures 1 and 2.

Coupling of Two Components: Assume that certain well defined properties of each component are given and that corresponding properties of the system are required. An example of such a property may be polynomials expressed as a function of frequency which specifies the natural vibration or eigenfunctions of a multiterminal system.

To illustrate the coupling procedure, assume that the property of the system to be investigated is the propagation of natural vibration frequencies from component specification to the desired system response. Therefore it is essential to know how any two components combine.

A systematic listing for combining components N and M together to form a system J is given in Table 1. The rules shown are 4-tuples (j,n,m,k) which result in a unique code to characterize coupling according to the number of terminals involved with components J, N, and M, and the resulting internal node k. To illustrate, consider rule 3331. This means that the three terminal system J (123) is formed by combining N (124) and M (234), and causing one internal connection to result. For two-terminal components only two types of coupling exist: (a) parallel: n = 2, m = 2, j = 2, k = 0, and (b) series: n = 2,

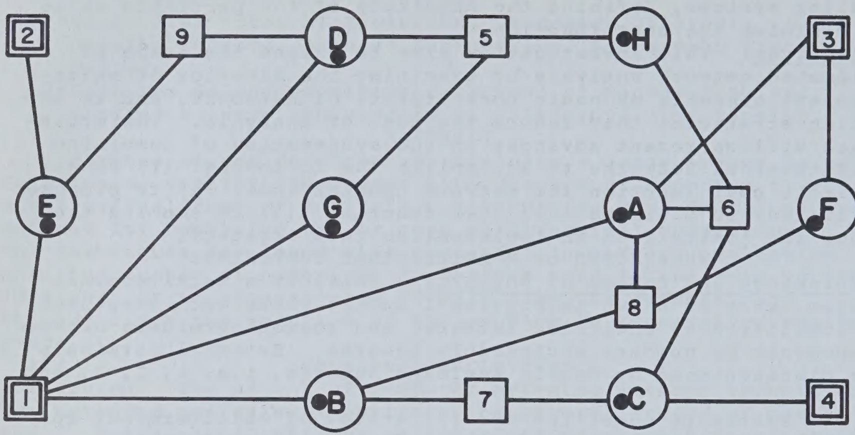

2
9
D
5
H
3
E
G
A
6
F
8
1
B
7
C
4

Figure 1. Multiterminal System / Z(1234) = ABCDEFGH

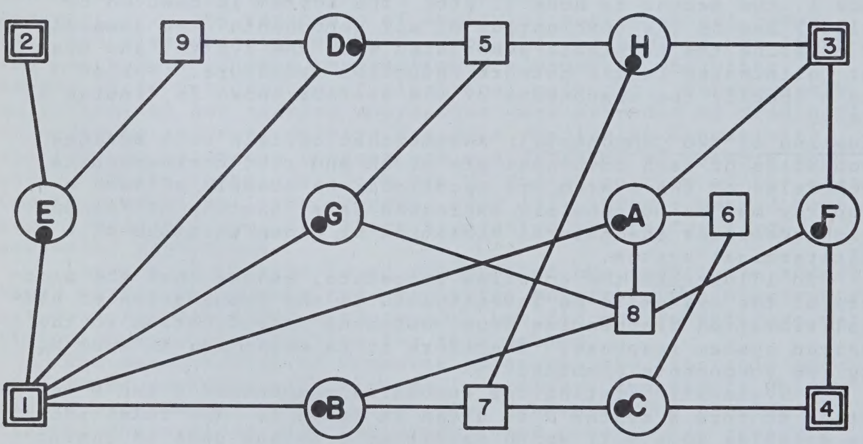

2
9
D
5
H
3
E
G
A
6
F
8
1
B
7
C
4

Figure 2. Additional Interconnections to System / Z(1234) = ABCDEFGH

$m = 2$, $j = 2$, $k = 1$.

For each procedure for combining components an algorithm must be developed. For two-terminal series-parallel networks only two algorithms are needed; for three terminal networks, ten rules are needed, and so on.

ELIMINATION STRATEGIES

Two elimination strategies are described: nearest neighbor and node elimination.

Nearest Neighbor: The nearest neighbor strategy consists of the successive steps: (1) Candidate selection: eligible components for combining are selected from neighboring components (that is nodes in common) or groups of components. (2) Cost function assignment: the price paid for combining two or more components is established by assessing the complexity of the resultant network. (3) Priority definition: guidelines for the sequence of eliminations are set but with the aim to reduce the overall cost function.

For large systems it appears difficult to list all possible sequences of combining components. A reasonable compromise consists of listing any two neighboring components as candidates. For this case the cost function E is the difference between the number of terminals of the resultant cluster and the number of terminals of the largest component.

Priority is given to candidates in the following sequence: (1) Select the minimum value of the cost function, E, which will eliminate a node, (2) Other factors being equal, select the lowest rule number (which actually means fewer terminals result), (3) If all previous tests are equal, select next candidate alphabetically.

Illustrative Examples: Tables 3 and 4 summarize the steps taken to reduce the system shown in Figure 1 by implementing the nearest neighbor strategy. The top part of Table 4 lists all possible candidates of two neighboring nodes. For each candidate the resulting terminal connections are given. The N column records the nodes that could be eliminated by combination. The E column records the cost incurred by combining components. Each + sign refers to the additional terminal that results in combining the two components over the largest number of terminals of the individual candidates. For example, the candidate AB can be combined using rule 5430 (which is not shown in Table 1, but will appear in a future paper), and since A has four terminals, the larger number of terminals of A and B, and since the resulting combination has five terminals, a + sign is inserted in the E column.

Once all the cost functions have been computed, the above selection procedure is then applied. For this example the DE combination which is defined as I is selected since E = 0 and a node is eliminated. All remaining combinations on the list which include DE are removed from further consideration. Such candidates are marked with an I followed by an asterisk in the selection column, e.g. I*.

The procedure is repeated with the newly formed five terminal component I, adding combined components J through N, until Z the over system function is reached. Table 3 summarizes the elimination sequence starting with the original elements and recording the nodes that are eliminated in the process. The letter entries are designations assigned to each combination carried through each step.

TABLE 1
ADMISSIBLE NETWORK COMBINATIONS

Rule	(j	n	m	k)	J(i...j)
2220	2	2	2	0	N(12) M(12)
2221				1	N(13) M(33)
2321	2	3	2	1	N(123) M(23)
2331	2	3	3	1	N(123) M(123)
2332				2	N(134) M(234)
2422	2	4	2	2	N(1234) M(34)
2432	2	4	3	2	N(1234) M(134)
2442	2	4	4	2	N(1234) M(1234)
2443				3	N(1345) M(2345)
3220	3	2	2	0	N(13) M(23)
3320	3	3	2	0	N(123) M(12)
3321				1	N(124) M(43)
3330	3	3	3	0	N(123) M(123)
3331				1	N(124) M(234)
3421	3	4	2	1	N(1234) M(41)
3431	3	4	3	1	N(1234) M(124)
3432				2	N(1245) M(345)
3441	3	4	4	1	N(1234) M(1234)
3442				2	N(1245) M(1345)
4320	4	3	2	0	N(123) M(14)
4330	4	3	3	0	N(123) M(124)
4331				1	N(125) M(345)
4420	4	4	2	0	N(1234) M(12)
4421				1	N(1235) M(45)
4430	4	4	3	0	N(1234) M(123)
4431				1	N(1235) M(145)
4440	4	4	4	0	N(1234) M(1234)
4441				1	N(1235) M(1245)
4442				2	N(1256) M(3456)

j = No. of external terminals of combined components
n = No. of external terminals of component N
m = No. of external terminals of component M
k = No. of internal terminals of combined components

TABLE 2
NON-REDUNDANT EIGENFUNCTIONS P(J) IN A J-TERMINAL NETWORK

J	2	3	4	5	6
P(J)	2	5	15	52	203

TABLE 3: COMPONENTS AND NEAREST NEIGHBOR ELIMINATION STRATEGY FOR FIGURE 1

Topology of Components	Elimination Sequence						
Z(1234)	9	0	5	0	8	7	6
A(1368)				L	M	N	Z
B(187)					M	N	Z
C(764)						N	Z
D(195)	I	J	K				Z
E(129)	I	J	K				Z
F(83)				L	M	N	Z
G(15)		J	K				Z
H(56)			K				Z

TABLE 4: NEAREST NEIGHBOR STRATEGY FOR FIGURE 1

Candidates		N	E	Rule	Selection
AB	13687	0	+	5430	L*
AC	134678	0	++	6430	L*
AF	1368	0	0	4420	L
AH	15368	0	+	5420	K*
BC	1864	7	+	4331	M*
BF	1387	0	+	4320	L*
CH	5647	0	+	4320	K*
DE	125	9	0	3331	I
DG	195	0	0	3320	I*
DH	1569	0	0	4320	I*
GH	156	0	+	3220	J*
GI	125	0	0	3320	J
HI	1265	0	+	4320	J*
HJ	126	5	0	3321	K
AK	12368	0	+	5430	L*
CK	12647	0	+	5430	N*
BL	1367	8	0	4431	M
CL	136478	0	++	6430	M*
KL	12638	0	+	5430	M*
CM	1364	7	0	4431	N
KM	12367	0	+	5430	N*
KN	1234	6	0	4431	Z

TABLE 5: COMPONENT AND NEAREST NEIGHBOR ELIMINATION STRATEGY FOR FIGURE 2

Topology of Components	Elimination Strategy						
Z(1234)	9	6	0	7	0	8	5
A(1368)		J	K	L	M	N	Z
B(187)			K	L	M	N	Z
C(764)		J	K	L	M	N	Z
D(519)	I						Z
E(129)	I						Z
F(834)					M	N	Z
G(158)						N	Z
H(753)				L	M	N	Z

TABLE 6: NEAREST NEIGHBOR STRATEGY FOR FIGURE 2

Candidates	N	E	Rule	Selection
AB 13687	0	+	5430	J*
AC 13478	6	+	5431	J
AF 13468	0	+	5430	J*
AG 15368	0	+	5430	J*
BC 18764	0	++	5330	J*
BF 18347	0	++	5330	K*
BG 1587	0	+	4330	K*
BH 15387	0	++	5330	K*
CH 53467	0	++	5330	J*
DE 125	9	0	3331	I
DG 1958	0	+	4330	I*
DH 19537	0	++	5330	I*
FG 15834	0	++	5330	M*
GH 15378	0	++	5330	L*
GI 1258	0	+	4330	N*
HI 12375	0	++	5330	L*
BJ 13478	0	0	5530	K
FJ 13478	0	0	5530	K*
GJ 153478	0	+	6530	K*
HJ 153478	0	+	6530	K*
FK 17348	0	0	5530	L*
GK 157348	0	+	6530	L*
HK 15348	7	0	5531	L
FL 15348	0	0	5530	M
GL 15348	0	0	5530	M*
IL 125348	0	+	6530	M*
GM 1534	8	-	4530	N
IM 125348	0	+	6530	N*
IN 1234	5	0	4430	Z

TABLE 7: PROCEDURE FOR ELIMINATING NODES

Node	Candidates		E	Combined Component
9	DE	125	0	I
8	ABFG	13457	+	
7	BCH	134568	+++	
6	AC	13478	+	
5	DGH	13789	++	
8	ABFG	13457	+	J
7	BCH	134568	+++	
6	AC	13478	+	
5	GHI	12378	++	
8	BFGJ	134578	+	M
7	BHJ	13458	0	
5	GHI	12378	++	
8	GM	1345	+	N
5	GI	1258	+	
5	IN	1234	0	Z

The nearest neighbor strategy was also applied to the system illustrated in Figure 2. The results are shown in Figures 5 and 6. Note the complexity that occurs when only small changes are added to Figure 1. Combination rules produced six terminal networks which first of all have to be established (and stored in the computer) and secondly could become expensive to compute if no alternative candidate occurs in the selection process. Another strategy could be applied.

Node Elimination: The node elimination strategy consists of procedures to identify all components common to the internal nodes, and to determine the number of terminals that would result by combining all common components and which would eliminate a node. The cost function, for this case, is defined similarly to the nearest neighbor strategy except that the largest number of terminals of the combining components is compared with the number of terminals of the combined cluster of components.

The priority is given to candidates in the following sequence: (1) Select the minimum value of E which will eliminate a node, (2) Other factors being equal, select the smallest number of terminals of the resulting cluster, (3) Other factors being equal, select the smallest number of components in the resultant cluster.

The node elimination procedure is given in Table 7 for the system illustrated in Figure 2. In the first selection procedure all internal nodes 5 through 9 are taken as candidates to determine the minimum cost of combination. For the first selection the combination of DE is of lowest cost, and is defined as I. The process is repeated for nodes 5 through 8 with I treated as a newly formed component. The AC combination is selected over ABEG since a smaller number of components are involved. The new combination is defined as J. The procedures continue until all nodes are eliminated resulting in the final representation of Z.

The strategies defined consist of step by step elimination. Since one susbystem per step is formed, the number of steps equals the number of components. Both strategies are surprisingly simple once they are formulated and hence easy to program. The nearest neighbor in simpler and adequate for systems with relatively few interconnections. The node elimination strategy requires more computational effort but is always at least as efficient as the nearest neighbor strategy.

CONCLUSION

Perhaps the most important conclusion is the realization that the network combination problem is not cost effective without a strategy. Computational procedures lacking any strategy are orders of magnitude more costly than procedures obtained by a relatively simple strategy. On the other hand, substantially improved strategies lower the cost only slightly if compared with more sophisticated strategies. Improved strategies have significant advantages only in highly interconnected systems.

Factors for defining cost effectiveness are categorized in Table 8 and form the topics of impeding investigations.

REFERENCES

1. Kron, G., "A Set of Principles to Interconnect the Solutions of Physical Systems", J. Applied Physics, V. 24, 1953, pp. 965-980.
2. Branin, F. H., "Kron's Method of Tearing and Its Applications", Proceedings Second Midwest Symposium on Circuit

Theory, Michigan State University, Dec. 1956.
3. "1620 Electronic Circuit Analysis Programs (ECAP)", IBM Corporation Data Processing Division, White Plains, N. Y. Application Program. 1620-EE-02X.
4. Murray-Lasso, M. A., "Analysis of Linear Integrated Circuits by Digital Computer Using Black Box Techniques", Chapter 4 in Computer-Aided Integrated Circuit Design, edited by G. Herskowitz, McGraw-Hill, 1968.
5. Kernighan, Brian, W., "Optimal Sequential Partitions of Graphs", J. of the Association for Computing Machinery, V. 18, No. 1, January 1971, pp. 34-40.
6. Kral, J., "To the Problem of Segmentation of a Program", Information Processing Machines, Research Institute for Mathematical Machines, Prague, 1965, pp. 140-149.
7. Berztiss, A. T., "A Note on Segmentation of Computer Programs", Information Control 12, 1, (January 1968), p. 21-22.
8. Dearnley, F. H., and Newell, G. B., "Automation Segmentation of Programs for a Two Level Store Computer", Computation Journal, 7, 3, (October 1964), pp. 185-187.

TABLE 8: COST EFFECTIVENESS FACTORS

Effectiveness Criteria	Parametric Representation	Algorithmic Implementation
Strategy	(a) elements (b) tie breaker (c) sequence	Procedural operations Single versus mixed strategies Fixed versus objective strategies
Cost	(a) elements (b) weights (c) trade-offs	Machine organizations-hardware Programming sophistication software Empirical field experience feedback
Topology	(a) N $\geq$ 3 terminals (b) number of elements (c) interconnect	Free concepts versus deviations techniques Depth of detail in modeling Physical partitioning

GENERALIZED INVERSES, DUALS, DUAL ADJOINTS AND APPLICATIONS

M.N.S. SWAMY, C. BHUSHAN and B.B. BHATTACHARYYA
Department of Electrical Engineering
Sir George Williams University, Montreal, Canada

ABSTRACT

For linear time invariant (LTI) networks, Mitra et. al.[1] have defined generalized inverse (GI) and generalized dual (GD) networks in the frequency domain. They define the GI of an n-port network N to be a network N_I for which the impedance matrix [z] is f(s) times the admittance matrix [y] of N. Depending on the nature of f(s), they also define resistive, inductive and capacitive inverses. They then define the GD of a network N consisting of one-port elements and controlled sources to be N_D for which the topology is the dual of that of N and each element is the GI of the corresponding element of N. They also show that the GD is always a GI, while the converse is not true in general; further they consider the application of capacitive inverses and duals in active lumped RC-network design.

In this paper, we first extend the definition of the GI for a general n-port consisting of linear/nonlinear, time varying/time invariant, lumped/distributed elements. Let V_i and I_i (i=1,2,...n) be the port voltages and currents of such a network N. Then we shall define N_I to be its GI if its port voltages V_i' and currents I_i' are given by: $V_i' = \lambda I_i$, $I_i' = \lambda' V_i$, where λ and λ' are linear operators. It is seen that in the case of LTI networks, this definition reduces, in the frequency domain, to that of Mitra et.al.[1]. We then define planarity for a general network N consisting as sub-networks one-ports and three terminal two-ports, and give a simple method of obtaining the dual topology of such a planar network. We then define the GD of N as that network N_D for which the topology is the dual of that of N and each internal one-port or three terminal two-port is the GI of the corresponding one in N. It can be shown that such a GD is also a GI while the converse is not true in general. It is shown that in the case of LTI networks, if each element in N_D is the GI of the corresponding element in N with respect to f(s), then N_D itself is the GD of N with respect to f(s). The GD's of different active elements and distributed elements have been obtained. For example, the capacitive duals of a VNIC, a VVT and an RC-tapered line are respectively a CNIC, a CCT and the dual RC-line (as defined in [2]), while the resistive duals of a CVT, a gyrator and a tapered lossless line are a VCT, a gyrator and the dual lossless line [2]. The capacitive duals can be used to get alternate RC lumped/distributed structures for realizing a driving point function (DPF), or for synthesizing a current transfer function (CTF), T(s), from a known structure realizing T(s) as a voltage transfer function (VTF). For example, the two Thomas' structures [3] realizing an arbitrary CTF are the capacitive duals of those of Yanagisawa [3] for realizing an arbitrary VTF. It is noted that the sensitivities of the corresponding network functions of N and N_D are the same with respect to corresponding network elements. It should also be pointed out that the GD in conjunction with the transpose [4] or adjoint [5] network may be used (i) to obtain three alternate structures for a DPF realization from a known realization for the given DPF, (ii) to obtain an alternate structure

for realizing a VTF (or CTF) starting with a known realization for the VTF (or CTF), and (iii) to obtain two structures for realizing a CTF (or VTF) from a known realization for a VTF (or CTF).

The second part of the paper is concerned with deriving a "Tellegen's theorem" for networks N and N' having dual topologies, in the form

$$\sum_b (\lambda\ i_b\ \lambda'\ i'_b \mp \lambda\ v_b\ \lambda'\ v'_b) = 0$$

where λ and λ' are linear operators, i_b and v_b are the currents and voltages of the branches and ports of N, while i'_b and v'_b are the currents and voltages of the corresponding branches and ports of N'. Using this Tellegen's theorem for dual topologies and following the procedure of Director and Rohrer [5], we can define a "generalized dual adjoint network" (GDA) and find the relations between the branch voltages and currents of the different network elements in the GDA. It is shown that if N_D, N_A and N' represent the Gd, the adjoint and the GDA of a network N then,

$$N' = (N_D)_A = (N_A)_D$$

Also, expressions for sensitivity components of "network functions" of N with respect to its different parameters may be obtained in terms of the branch voltages and currents of N and N' under suitable excitations. Thus, two network analyses are sufficient to determine the sensitivity of any network function of a planar network N with respect to any of its parameters. Further, the GDA may directly be used to get an alternate realization for a VTF (or CTF) from a known realization for the VTF (or CTF); for example, the two Yanagisawa's structures for realizing a VTF are capacitive dual adjoints of each other, with a similar statement holding true for those of Thomas for realizing a CTF.

ACKNOWLEDGEMENT

This work was supported by National Research Council of Canada, Grant No. A-7739, and is based on a part of a doctoral thesis, to be submitted by C. Bhushan under the supervision of Prof. M.N.S. Swamy.

REFERENCES

[1] S.K. Mitra, N. Herbst and N. Declaris, "Generalised dual and inverse networks: An extended definition", IEEE international convention record 11 (Part 2) 1963, pp 76-82.
[2] B.B. Bhattacharyya and M.N.S. Swamy, "Dual distribution of solvable nonuniform lines", Proceedings IEEE, vol. 54, No. 12, 1966, pp 1979-1980.
[3] S.K. Mitra, "Analysis and synthesis of linear active networks", John Wiley and Sons, 1969.
[4] B.B. Bhattacharyya and M.N.S. Swamy, "Network transposition and its application in synthesis", IEEE Trans. 1971, CT-18, pp 394-397.
[5] S.W. Director and R.A. Rohrer, "The generalised adjoint network and network sensitivities", IEEE Trans., 1969, CT-16, pp 318-323.

STATE-SPACE INTERPRETATION OF CLASSICAL RESULTS IN NETWORK THEORY

M.LAL AND HARPREET SINGH
Department of Electronics and Communication Engrg.,
University of Roorkee, Roorkee. India.

ABSTRACT

It has been shown that it is possible to derive classical results in network theory in state-space terms either from state-models or from input-output characterization in s-domain. The well-known properties of LC, RC networks etc. have been derived in state-space terms. The state-space interpretation of classical Foster and Cauer methods has been given.

I. INTRODUCTION

With the acceptance of state-space technique as a powerful tool for network theory [1]-[7], it appears worthwhile to interpret the classical results of network analysis and synthesis in state-space terms. The state-space interpretation of poles, zeros [1], positive real functions [2], Foster method [3], Darlington method [4] and the methods for determining impedance matrix from given even part [5] has been done. The object of the present paper is to show that various other classical results in network analysis and synthesis can be interpreted in state-space terms from both the commonly used characterizations i.e. state-model characterization and input-output characterization. In particular, a few well-known properties of LC, RC networks etc. have been derived using state space ideas and interpretation of classical Foster and Cauer methods has been done in state-space terms as discussed below.

II. DERIVATION OF SOME OF THE NETWORK-PROPERTIES USING STATE-SPACE METHODS

Property No.1

Let $Z(s)$ be a lossless impedance matrix with $Z(\infty)=0$ and let $[A,B,C]$ be a minimal realization for $Z(s)$ in the sense that

$$Z(s) = C(sI - A)^{-1}B \tag{1}$$

and A has minimal dimension. Then there exists a symmetric positive definite matrix P such that

$$PA + A'P = 0 , \tag{2a}$$

$$PB = C', \tag{2b}$$

where superscript prime denotes matrix transposition.

Proof The property although proved in [2] from general expression of LC networks in s-domain, is proved here from state-model characterization of LC networks. The proposed method gives better insight because state models are the basic building blocks in modern network synthesis just as loop and node methods are in classical synthesis.

Consider the generalized state model of LC networks in topological entities in the form [3]

$$\dot{X} = AX + B_t U + B_o \dot{U} \, , \tag{3a}$$

$$Y = CX + D_t U + D_o \dot{U} \, , \tag{3b}$$

where X is n-vector, the state; U is m vector, the input, and Y is p-vector, the output given by

$$X = \begin{bmatrix} V_C \\ I_L \end{bmatrix}, \quad U = \begin{bmatrix} \overset{*}{V}_V \\ \overset{*}{I}_K \end{bmatrix}, \quad Y = \begin{bmatrix} \overset{*}{V}_K \\ \overset{*}{I}_V \end{bmatrix} \tag{4}$$

where $\overset{*}{V}_V = V_V$, $\overset{*}{V}_K$ and $\overset{*}{I}_V = -I_V$ and $\overset{*}{I}_K = -I_K$ represent the vectors of terminal variables for n-port LC network. A, B_t, B_o, C, D_t, D_o are nxn, nxm, nxm, pxn, pxm and pxm dimensional matrices and are given as

$$A = \begin{bmatrix} \mathcal{C}^{-1} & 0 \\ 0 & \mathcal{L}^{-1} \end{bmatrix} \begin{bmatrix} 0 & F'_{LC} \\ -F_{LC} & 0 \end{bmatrix}, \tag{5a}$$

$$B_t = \begin{bmatrix} \mathcal{C}^{-1} & 0 \\ 0 & \mathcal{L}^{-1} \end{bmatrix} \begin{bmatrix} 0 & -F'_{KC} \\ -F_{LV} & 0 \end{bmatrix}, \tag{5b}$$

$$B_o = \begin{bmatrix} \mathcal{C}^{-1} & 0 \\ 0 & \mathcal{L}^{-1} \end{bmatrix} \begin{bmatrix} -F'_{SC} C_1 F_{SV} & 0 \\ 0 & F_{L\Gamma} L_2 F'_{K\Gamma} \end{bmatrix} \tag{5c}$$

$$C = \begin{bmatrix} 0 & F'_{LV} + (F'_{SC} C_1 F_{SV})' \mathcal{C}^{-1} F'_{LC} \\ -F_{KC} + (F_{L\Gamma} L_2 F'_{K\Gamma})' \mathcal{L}^{-1} F_{LC} & 0 \end{bmatrix} \text{etc.,} \tag{5d}$$

where $\mathcal{C}$ and $\mathcal{L}$ (both symmetric and positive definite) are given by

$$\mathcal{C} = C_2 + F'_{SC} C_1 F_{SC} \, , \tag{6a}$$

$$\mathcal{L} = L_1 + F_{L\Gamma} L_2 F'_{L\Gamma} . \tag{6b}$$

Symbols C,L,V,I stand for capacitance, inductance, voltage source and current source respectively. Subscripts 1 and 2 denote links and tree branches respectively for the normal tree chosen. Further the subscripts S,L and K denote link elastance, inductance and current source and subscripts C,Γ and V denote capacitance, reciprocal inductance and voltage source respectively. The terminology used here is similar to Kuh and Rohrer [8].

The strategy lies in choosing a transformation [7]

$$X = X_1 + B_o\dot{U} \tag{7}$$

which when applied to the state model (3) gives the state model without source derivatives in the form

$$\dot{X}_1 = AX_1 + BU , \tag{8a}$$

$$Y = CX_1 + DU , \tag{8b}$$

where

$$B = B_t + AB_o$$

$$= \begin{bmatrix} \mathcal{C}^{-1} & 0 \\ 0 & \mathcal{L}^{-1} \end{bmatrix} \begin{bmatrix} 0 & -F'_{KC}+F'_{LC}\mathcal{L}^{-1}(F_{L\Gamma}\;L_2F'_{K\Gamma}) \\ -F_{LV}+F_{LC}\,\mathcal{C}^{-1}(F'_{SC}C_1F_{SV}) & 0 \end{bmatrix}, \tag{9}$$

Taking

$$P = \begin{bmatrix} \mathcal{C} & 0 \\ 0 & \mathcal{L} \end{bmatrix} \tag{10}$$

the property can be easily proved.

Property No.2

Poles and zeros of LC driving-point function lie on imaginary axis.

Proof A proper driving-point function z(s), i.e. vanishing at $s = \infty$, and having realization [A,b,c] can be written as [9]

$$z(s) = \frac{\det\begin{bmatrix} 0 & -c \\ b & sI-A \end{bmatrix}}{\det\begin{bmatrix} sI - A \end{bmatrix}} \; . \tag{11}$$

For LC case, if the given driving-point function is not proper, its reciprocal must be proper. Therefore no generality is lost in starting with a proper LC driving-point function.

Applying transformation

$$T = P^{-1/2} \tag{12}$$

we obtain transformed $\left[A, b, c\right]$ as $\left[P^{1/2}AP^{-1/2},\ P^{1/2}b, cP^{-1/2}\right]$ which when substituted in (2) shows that the new A matrix is skew and the new b equals c'. From (11) the property is thus self-evident.

Property No.3

Poles and zeros of LC driving-point function interlace.

Proof In a suitable neighbourhood of infinity, z(s) can be expanded in terms of Markov parameters [10] etc. in the form

$$z(s) = cbs^{-1} + cAbs^{-2} + cA^2bs^{-3} + \ldots + \ldots \tag{13}$$

For reactive networks, A, b, c can be transformed to a form such that new A is skew and new b equals new c' as indicated in property no.2. Therefore, from (13) we can prove that the slope of reactance versus frequency curve is always positive keeping in mind that, for the case under consideration, cc' is always positive, $cAc' = 0$ and $cA^2c' \leqslant 0$ etc.

Property No.4

Poles and zeros of RC driving-point function lie on negative real axis and interlace.

Proof By following the approach used in [3] or [8] and letting only a current source at the input, the state model for RC impedance function can be obtained in the form

$$\dot{X} = AX + bu, \tag{14a}$$

$$y = cX + du, \tag{14b}$$

which in topological entities is written as

$$\begin{aligned} \dot{V}_C &= \mathcal{C}^{-1}\left[-F'_{RC}\ \mathcal{R}^{-1}F_{RC}\right] V_C + \mathcal{C}^{-1}\left[F'_{KC} + F'_{RC}\ \mathcal{R}^{-1}F_{RG}R_2F'_{KG}\right]\overset{*}{i}_K, \\ \overset{*}{v}_K &= \left[-F_{KC} + F_{KG}\,\mathcal{G}^{-1}F'_{RG}G_1F_{RC}\right] V_C + F_{KG}\,\mathcal{G}^{-1}F'_{KG}\ \overset{*}{i}_K \end{aligned} \tag{15}$$

where subscripts R and G denote link resistance and tree-branch conductance respectively, $\overset{*}{v}_K$ and $\overset{*}{i}_K$ denote the voltage and the current output of the current source etc. and $\mathcal{R}$ and $\mathcal{G}$ (both positive definite) are given by

$$\mathcal{R} = R_1 + F_{RG}R_2F'_{RG} \, , \tag{16a}$$
$$\mathcal{G} = G_2 + F'_{RG}G_1F_{RG} \, . \tag{16b}$$

Applying the transformation

$$T = \mathcal{C}^{-1/2} \tag{17}$$

to state model (15), we obtain a state model in which the new A can be proved to be symmetric negative definite matrix and so has negative real eigen values and matrix

$$A^* \triangleq A - \frac{bc}{d} \tag{18}$$

giving zeros [9] comes to be symmetric negative semidefinite and thus giving zeros of the function as negative real. Application of the bordering property [11] proves the interlacing part. Similar results using the above ideas can be proved for RL case.

III. STATE-SPACE INTERPRETATION OF CLASSICAL SYNTHESIS METHODS

Foster Method (RC Networks)

The state-space interpretation of classical Foster method for driving-point LC functions has been discussed [3]. Similar interpretation can be given for RC case for which z(s) can be expressed as

$$z(s) = K_o + \frac{K_1}{s + \sigma_1} + \frac{K_2}{s + \sigma_2} + \dots + \frac{K_n}{s + \sigma_n} \, . \tag{19}$$

where σ's and K's are positive constants. A state model corresponding to z(s) can be written as

$$A = \begin{bmatrix} -\sigma_1 & & & \\ & -\sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix}, \qquad b = \begin{bmatrix} \sqrt{K_1} \\ \sqrt{K_2} \\ \vdots \\ \sqrt{K_n} \end{bmatrix}, \tag{20}$$

$$c = \begin{bmatrix} \sqrt{K_1} & \sqrt{K_2} & \cdots & \sqrt{K_n} \end{bmatrix}, \qquad d = \begin{bmatrix} K_o \end{bmatrix} .$$

Apply transformation

$$T = \operatorname{diag}\left[\frac{1}{\sqrt{K_1}} \, , \frac{1}{\sqrt{K_2}} \, , \dots , \frac{1}{\sqrt{K_n}} \right] . \tag{21}$$

Decompose the resulting state equations by comparing with the state model obtained from (15) for the class of RC networks having $F_{RG} = 0$ [8] and noting that $\mathcal{C}$ is symmetric positive definite matrix and entries of F'_{KC} etc. are to be ±1,0. Element values of capacitances can thus be found as

$$= \operatorname{diag}\left[\frac{1}{K_1}, \frac{1}{K_2}, \ldots, \frac{1}{K_n}\right] \qquad (22)$$

and the circuit matrix $\begin{bmatrix} F & I \end{bmatrix}$ is obtained as

$$\begin{bmatrix} F & I \end{bmatrix} = \begin{array}{c} \\ K \\ R_1 \\ R_2 \\ \vdots \\ R_n \end{array} \begin{array}{c} \begin{array}{ccccc} G & C_1 & C_2 & \ldots & C_n \end{array} \\ \left[\begin{array}{ccccc|ccccc} 1 & -1 & -1 & & -1 & 1 & & & & \\ 0 & 1 & 0 & \ldots & 0 & & 1 & & & \\ 0 & 0 & 1 & \ldots & 0 & & & 1 & & \\ \vdots & \vdots & \vdots & & \vdots & & & & \ddots & \\ 0 & 0 & 0 & & 1 & & & & & 1 \end{array}\right] \end{array}, \qquad (23)$$

which is realizable and results in classical Foster Form. (Fig.1)

Cauer Method (LC Networks)

Consider a proper driving-point reactance function

$$z_1(s) = \frac{a_1 s^{n-1} + a_3 s^{n-3} + \ldots + a_n s}{s^n + b_2 s^{n-2} + \ldots + b_n} . \qquad (24)$$

Write down the coefficients of numerator and denominator for determining Hurwitz coefficients in the usual tabular form as given below,

$$\begin{array}{llll} k_o = 1 & b_2 & \ldots & b_n \\ k_1 = a_1 & a_3 & \ldots & a_{n-1} \\ \vdots & \vdots & & \\ k_n & \cdot & & \end{array} \qquad (25)$$

where $k_o=1$, $k_1 = a_1$ etc. denote the entries in the first column of (25) and, obviously, are all positive for the case under consideration. It can be easily verified [7] that the state model for (24) can be written in the modified Schwarz form in terms of the entries in the first column of (25) as given below.

$$\dot{X} = \begin{bmatrix} 0 & -k_2/k_o & & & & & & \\ 1 & 0 & -k_3/k_1 & & & & & \\ & 1 & 0 & -k_4/k_2 & & & & \\ & & 1 & 0 & -k_5/k_3 & & & \\ & & & 1 & 0 & -k_6/k_4 & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & 1 & 0 & -k_n/k_{n-2} \\ & & & & & & 1 & 0 \end{bmatrix} X + \begin{bmatrix} 1/k_o \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} u$$

$$y = \begin{bmatrix} k_1 & 0 & 0 & 0 & 0 & 0 \cdots & 0 \end{bmatrix} X \tag{26}$$

Next, applying the transformation,

$$T = \mathrm{diag}\left[\frac{1}{k_1}, \frac{1}{k_2}, \frac{1}{k_3}, \frac{1}{k_4}, \cdots, \frac{1}{k_{n-1}}, \frac{1}{k_n}\right], \tag{27}$$

we obtain a state model, the matrix A of which is naturally again in a tridiagonal form. This state model is subjected to transformation[12]

$$T = \begin{array}{cccccccccc} 1 & 2 & 3 & \frac{n+1}{2} & \frac{n+3}{2} & & & n-1 & n & \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 4 \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & n-1 \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & n \end{array} \tag{28a}$$

when n is odd and

$$T = \begin{array}{cccccccccc} 1 & 2 & 3 & \frac{n}{2} & \frac{n}{2}+1 & & & n-1 & n & \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 4 \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & n-1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & n \end{array} \tag{28b}$$

when n is even. The resulting state model can be decomposed in the form

$$\begin{aligned} \dot{X} &= \Lambda A_1 X + \Lambda b_1 u , \\ y &= cX , \end{aligned} \tag{29}$$

where diagonal matrix Λ with positive entries is given by

$$\mathrm{diag}\left[\Lambda_C, \Lambda_L\right] = \mathrm{diag}\left[\frac{k_1}{k_o}, \frac{k_3}{k_2}, \frac{k_5}{k_4}, \ldots, \frac{k_n}{k_{n-1}}, \frac{k_2}{k_1}, \frac{k_4}{k_3}, \ldots, \frac{k_{n-1}}{k_{n-2}}\right] \quad (30)$$

Compare state model (29) with the corresponding state model for LC one port obtained from (3) by taking only a current source at the port and given as

$$\begin{bmatrix} \dot{V}_C \\ \dot{I}_L \end{bmatrix} = \begin{bmatrix} C_2^{-1} & \\ & L_1^{-1} \end{bmatrix} \begin{bmatrix} 0 & F'_{LC} \\ -F_{LC} & 0 \end{bmatrix} \begin{bmatrix} V_C \\ I_L \end{bmatrix} + \begin{bmatrix} C_2^{-1} & \\ & L_1^{-1} \end{bmatrix} \begin{bmatrix} -F'_{KC} \\ 0 \end{bmatrix} \overset{*}{i}_K$$

$$\overset{*}{v}_K = \begin{bmatrix} -F_{KC} & 0 \end{bmatrix} \begin{bmatrix} V_C \\ I_L \end{bmatrix} . \quad (31)$$

Element values of capacitances and inductances can thus be found and are given by (30) and the circuit matrix $[F \quad I]$ is obtained as

$$[F \quad I] = \begin{array}{c|ccccccccc|} & C_1 & C_3 & C_5 & \cdots & C_n & & & & & \\ L_2 & -1 & 1 & & & & 1 & & & & \\ L_4 & & -1 & 1 & & & & 1 & & & \\ \vdots & & & \ddots & \ddots & & & & \ddots & & \\ L_{n-1} & & & -1 & & 1 & & & & 1 & \\ K & -1 & 0 & 0 & \cdots & 0 & & & & & 1 \end{array} \quad (32)$$

which is realizable and yields the well-known Cauer form.(Fig.2)

ACKNOWLEDGMENT

The authors wish to express their gratitude to Prof. A.K. Kamal for advice and encouragement.

REFERENCES

[1] E.S. Kuh, "State variable and feedback theory"IEEE Trans. on Circuit Theory, Vol.CT-16, No.1, pp.23-26, Feb.1969.

[2] B.D.O. Anderson, "A system theory criterion for positive real matrices", SIAM Journal on Control, Vol.5, No.2, pp.171-182, 1967.

[3] R.Yarlagadda and Y.Tokad, "Synthesis of LC networks-a state-model approach", Proc. of IEE, Vol.113, No.6, pp.975-981, June 1966.

[4] B.D.O. Anderson and R.W. Brocket, "A multiport state-space Darlington synthesis", IEEE Trans. on Circuit Theory, Vol.CT-14,No.3,pp.336-337, Sept.1967.

[5] M.Lal and Harpreet Singh, "Determination of impedance matrix Z(s), given Z(s) + Z'(-s), using state-variable technique", To appear in Jan. 1972 issue of IEEE Trans.on Circuit Theory.

[6] M.Lal and Harpreet Singh, "Computer algorithm for the synthesis of LC networks- A state-model approach", Presented in IEEE International Symposium on Electric Network Theory held from September 6 to September 10, 1971.

[7] Harpreet Singh, "State-space approach to Network Synthesis", Ph.D. Thesis, University of Roorkee,Roorkee, 1971.

[8] E.S. Kuh and R.A. Rohrer, The state variable approach to network analysis, Proc. IEEE, Vol.53, No.7, pp.672-686, July 1965.

[9] I.W. Sandberg and H.C. So, "Two sets of eigen value approach", IEEE. Trans. on Circuit Theory, Vol.CT-16,No.4, p.509, Nov. 1969.'

[10] B.L.Ho and R.E. Kalman, "Effective construction of linear state variable models from input/output data", Proceedings of the Third Allerton Conference on Circuit and System Theory, pp.449-459, 1965.

[11] A.S. Householder "The Theory of Matrices in Numerical Analysis", Wallham Mass Blaisdell, p.90,1964.

[12] R.Yarlagadda, "An application of tridiagonal matrix to network synthesis", SIAM Journal on Applied Mathematics, Vol.16, No.6, pp.1146-1162, Nov.1968.

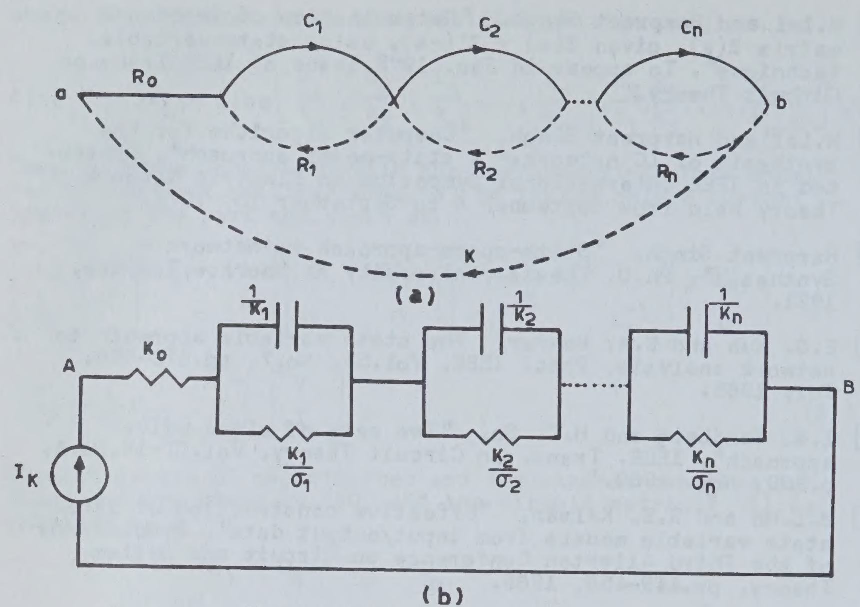

C1
C2
Cn
R0
a
b
R1
R2
Rn
K
(a)
1/K1
1/K2
1/Kn
K0
A
B
IK
K1/σ1
K2/σ2
Kn/σn
(b)

Figure 1. Foster form
(a) Realization of circuit matrix (23)
(b) Realization of $z(s)$ (19)

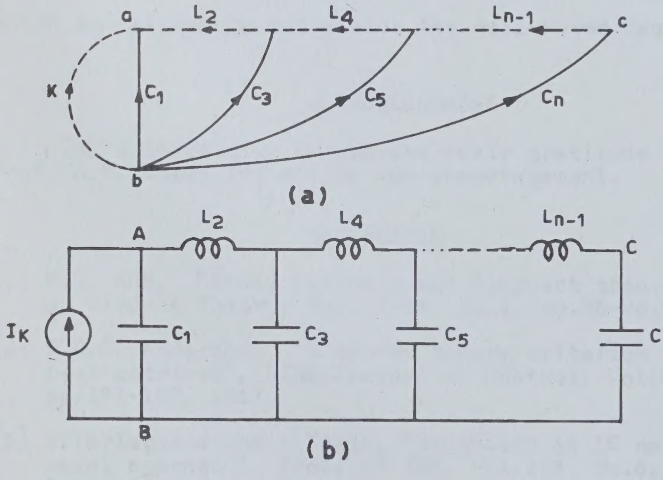

a
L2
L4
Ln-1
c
K
C1
C3
C5
Cn
b
(a)
L2
L4
Ln-1
A
C
IK
C1
C3
C5
Cn
B
(b)

Figure 2. Cauer form
(a) Realization of circuit matrix (32)
(b) Realization of $z_1(s)$ (24)

OPTIMUM SENSITIVITY DESIGN FOR A CLASS OF ACTIVE FILTERS

WILLIAM S. WAGNER
University of Cincinnati
Cincinnati, Ohio

ABSTRACT

This paper describes a design method which achieves optimum quality factor sensitivity and optimum transfer function sensitivity - both sensitivities being with respect to the positive gain active device - for a certain class of low gain active filters.

PROCEDURE

It is well-known[1] that the optimum Q sensitivity for the class of active filters described in Fig. 1 is as follows:

$$S_K^Q = 2Q - 1 \tag{1}$$

It can be shown that[2]

$$S_K^Q = Q\frac{b}{d^{.5}} - 1 \tag{2}$$

and also that,

$$S_K^{G(j\omega_o)} = Q\frac{b}{d^{.5}} \tag{3}$$

Thus, in order to optimize these sensitivities for a desired value of Q it is necessary to minimize the quantity $b/d^{.5}$, which is a function of G_1, G_2, G_3, C_1 and C_2. The minimization is achieved by taking partial derivatives of $b/d^{.5}$ with respect to the various components, equating the results to zero, and solving the ensuing equations for the relationship among the components.

For the low pass filter and the high pass filter this procedure leads to the following result:

$$\frac{C_1}{C_2} = \frac{G_1+G_2}{G_2} \gg 1 \tag{4}$$

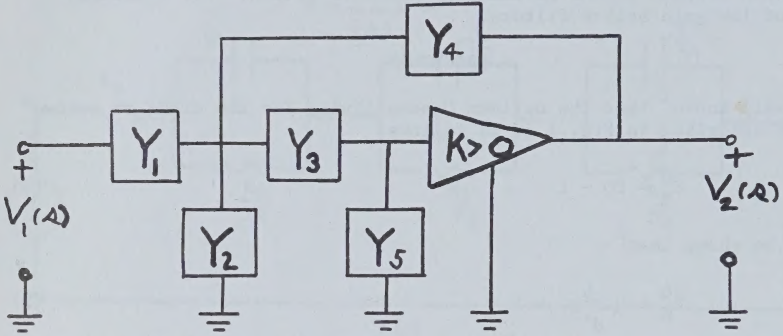
Similarly, the bandpass filter sensitivities are minimized by letting

$$\frac{C_1}{C_2} = \frac{G_1+G_2+G_3}{G_3} \gg 1 \tag{5}$$

REFERENCES

1. L. P. Huelsman, "Theory and Design of Active R.C. Circuits" McGraw-Hill Book Company, New York, N.Y., 1968, Chapter 3.

2. W. S. Wagner, "Sensitivity of Low Gain Active Filters", Proceedings of the 14th Midwest Symposium on Circuit Theory, Denver, Colorado, May 5-7, 1971.



$$G(s) = \frac{V_2(s)}{V_1(s)} = \frac{K a s^n}{s^2 + (b - cK)s + d}$$

n	Y_1	Y_2	Y_3	Y_4	Y_5	a	b	c	d
0 L.P.	G_1	0	G_2	C_1	C_2	$\frac{G_1G_2}{C_1C_2}$	$\frac{G_1}{C_1}+\frac{G_2}{C_1}+\frac{G_2}{C_2}$	$\frac{G_2}{C_2}$	$\frac{G_1G_2}{C_1C_2}$
1 B.P.	G_1	C_1	C_2	G_2	G_3	$\frac{G_1}{C_1}$	$\frac{G_1}{C_1}+\frac{G_2}{C_1}+\frac{G_3}{C_1}+\frac{G_3}{C_2}$	$\frac{G_2}{C_1}$	$\frac{G_2(G_1+G_2)}{C_1C_2}$
2 H.P.	C_1	0	C_2	G_1	G_2	1	$\frac{G_1}{C_1}+\frac{G_2}{C_1}+\frac{G_2}{C_2}$	$\frac{G_1}{C_1}$	$\frac{G_1G_2}{C_1C_2}$

Fig. 1 The Active Filter

MINIMIZATION OF THE QUALITY FACTOR SENSITIVITY

C.S. Phan and H.K. Kim
Department of Electrical Engineering
The University of Manitoba
Winnipeg, Canada

In the frequency-selective network, the quality factor Q determines the shape of the resonance. The effect of parameter variation on the resonance is to be minimized through Q-sensitivity study.

The Q-sensitivity is related [1], [2] to the pole sensitivity as

$$S_x^Q = -2Q \operatorname{Im}\{\frac{S_x^P}{p}\} \tag{1}$$

where p represents a pole and x is the parameter of interest. In this paper we will study the decomposition technique to minimize specifically the Q-sensitivity. Let T(s) be a transfer function with a conjugated pair of high Q poles as follows

$$T(s) = \frac{N(s)}{D(s)} = \frac{Hs}{s^2 + 2\sigma s + (\sigma^2 + \omega^2)} = \frac{Hs}{s^2 + 2\sigma s + \omega_o^2} \tag{2}$$

where $\sigma<<\omega$, and therefore, the quality factor Q is written as

$$Q = \frac{\omega}{2\sigma} \approx \frac{\omega_o}{2\sigma} \tag{3}$$

Let us now decompose D(s) in the most general form as follows where x is the parameter of our interest.

$$D(s) = D_1(s) + xD_2(s) = a_o(s-a_1)(s-a_2) + x(\alpha s-b_1)(\beta s-b_2) \tag{4}$$

Using (1) we derive

$$S_x^Q = \frac{a_o(-a_1-a_2)}{2\sigma} \tag{5}$$

Since $D_1(s) = D(s) - xD_2(s)$, it follows that

$$a_o(s-a_1)(s-a_2) = (1-x\alpha\beta)s^2 + [2\sigma+x(\alpha b_2+\beta b_1)]s + (\omega_o^2 -xb_1b_2) \tag{6}$$

and we express a_o, a_1, a_2 and the Q-sensitivity in terms of α,β,b_1b_2

$$a_o = 1 - x\alpha\beta > 0 \tag{7}$$

$$a_1,a_2 = \frac{-F \pm \sqrt{F^2 - G}}{2a_o}$$

$$2\sigma S_x^Q = a_o(-a_1-a_2) = F \tag{9}$$

where

$$F(\alpha,\beta,b_1,b_2) = 2\sigma + x(\alpha b_2 + \beta b_1) \tag{10}$$

$$G(\alpha,\beta,b_1,b_2) = 4(1-x\alpha\beta)\ (\omega_o^2 - xb_1b_2) \tag{11}$$

In order for the decomposition to be amenable to RC realization, it is required that both a_1 and a_2 be negative real. Thus, the Q-sensitivity minimization is reduced to finding the smallest F under the constraint $F^2 - G \geq 0$. Since $\sigma<<\omega_o$ the constraint rules out any combinations of α,β,b_1 and b_2 which yield $F^2 - G = 4(\sigma^2 - \omega_o^2)$.

A. Decomposition of D(s) ---- The Difference Case

In the difference decomposition, $x = -k$ $(k>0)$, and

$$F = 2\sigma - k(\alpha b_2 + \beta b_1) > 0 \tag{12a}$$

$$G = 4(1 + k\alpha\beta)\,(\omega_o^2 + kb_1b_2) > 0 \tag{12b}$$

F can be minimized by letting $F^2 - G = 0$ for smallest G which is found by $\alpha = b_2 = 0$ (or $\beta = b_1 = 0$) as follows

$$(F^2 - G)\Big|_{\alpha=b_2=0} = (2\sigma - k\beta b_1)^2 - 4\omega_o^2 = 0 \tag{13a}$$

or

$$k\beta b_1 = 2(\sigma - \omega_o) \tag{13b}$$

With $\alpha = b_2 = 0$, and (13b), we obtain

$$a_o = 1 \tag{14}$$

$$a_1 = a_2 = \frac{-F_{min}}{2a_o} = -\omega_o \tag{15}$$

Thus, it follows that

$$D_1(s) = a_o(s-a_1)\,(s-a_2) = (s + \omega_o)^2 \tag{16}$$

$$D_2(s) = (\alpha s-b_1)\,(\beta s-b_2)s = \frac{2(\omega_o - \sigma)s}{k} \tag{17}$$

$$S_k^Q = \frac{a_o(-a_1-a_2)}{2\sigma} = 2Q \tag{18}$$

B. Decomposition of D(s) ---- The Sum Case

In the sum case, $x = k$ $(k > 0)$, and

$$F = 2\sigma + k(\alpha b_2 + \beta b_1) \tag{19}$$

$$G = 4(1 - k\alpha\beta)\,(\omega_o^2 - kb_1b_2) \tag{20}$$

F is minimized by letting $F^2 - G = 0$ for smallest G. Since $kb_1b_2 > 0$, and $k(\alpha b_2 + \beta b_1) < 0$, G and F may vanish simultaneously under

$$kb_1b_2 = \omega_o^2 \tag{21}$$

$$k(\alpha b_2 + \beta b_1) = -2\sigma \tag{22}$$

yielding $F_{min} = G_{min} = 0$.

With $\alpha = \beta = \sigma/(\sqrt{k}\omega_o)$ and $b_1 = b_2 = -\omega_o/\sqrt{k}$, it follows that

$$a_o = 1 - k\alpha\beta = 1 - \sigma^2/\omega_o^2 > 0 \tag{23}$$

$$a_1 = a_2 = F_{min}/2a_o = 0 \tag{24}$$

Thus, we obtain

$$D_1(s) = (1 - \frac{\sigma^2}{\omega_o^2})s^2 \tag{25}$$

$$D_2(s) = \frac{\sigma^2}{k\omega_o^2}\,(s + \frac{\omega_o^2}{\sigma})^2 \tag{26}$$

$$S_x^Q = 0 \tag{27}$$

REFERENCES

[1] S. K. Mitra, Analysis and Synthesis of Linear Active Networks, New York: John Wiley, 1969, pp. 195-196.

[2] G. S. Moschytz, "A note on pole, frequency, and Q sensitivity," IEEE J. Solid-State Circuits, vol. SC-6, Aug. 1971, pp. 267-269.

A NEW LOOK AT PERIODIC CODES

DANIEL J. COSTELLO, JR.
Department of Electrical Engineering
Illinois Institute of Technology
Chicago, Illinois 60616

ABSTRACT

Many of the most advantageous properties attributed to convolutional codes can be proved only for the class of time-varying codes. This has led to considerable speculation that good specific time-varying codes can be found which outperform the best fixed codes. In this paper we introduce a new class of simply instrumented periodic time-varying codes and investigate their asymptotic properties. Specifically, we show that:

(1) They meet the Gilbert bound on minimim distance;

(2) They meet a bound on free distance equal to twice the Gilbert bound for rates $R > 3/8$.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation (NSF Grant GK-5265).

A NEW CLASS OF CODES

We wish to consider a class of binary non-systematic periodic time-varying convolutional codes with period $T = 2m+1$, where m is the memory order of the encoder. For the rate $R = 1/2$ case, the encoder can be realized with the simple circuit shown in figure 1. Each time the center register shifts, the other two registers shift twice. For other rates, the encoder is a straightforward generalization of the one shown in figure 1. Clearly the encoder has period $T = 2m+1$ since the contents of the top and bottom registers are exactly the same after $2m+1$ double shifts. We now proceed to obtain lower bounds on both minimum distance and free distance for this class of codes. The results will be proved for the important case of $R = 1/2$, and the straightforward generalization to other rates will be stated as corollaries. In the $R = 1/2$ case, the encoding equations can be written as:

$$\begin{aligned}
y_0 &= x_0 g_0 \\
y_1 &= x_1 g_2 + x_0 g_1 \\
y_2 &= x_2 g_4 + x_1 g_3 + x_0 g_2 \\
&\vdots \\
y_{m-2} &= x_{m-2} g_{2m-4} + x_{m-3} g_{2m-5} + \dots + x_0 g_{m-2} \\
y_{m-1} &= x_{m-1} g_{2m-2} + x_{m-2} g_{2m-3} + \dots + x_1 g_m + x_0 g_{m-1} \\
y_m &= x_m g_{2m} + x_{m-1} g_{2m-1} + \dots + x_2 g_{m+2} + x_1 g_{m+1} + x_0 g_m \\
&\vdots
\end{aligned} \tag{1}$$

A superscript $^{(1)}$ on all y's and g's refers to the equations for the first transmitted sequence $\underline{y}^{(1)} = [y_0^{(1)}\ y_1^{(1)}\ y_2^{(1)} \dots]$ and a superscript $^{(2)}$ refers to the second transmitted sequence $\underline{y}^{(2)} = [y_0^{(2)}\ y_1^{(2)}\ y_2^{(2)} \dots]$. The data or information sequence is denoted $\underline{x} = [x_0\ x_1\ x_2 \dots]$. Clearly the quality of the code is determined by the two generator sequences $\underline{g}^{(1)} = [g_0^{(1)}\ g_1^{(1)}\ g_2^{(1)} \dots g_{2m}^{(1)}]$ and $\underline{g}^{(2)} = [g_0^{(2)}\ g_1^{(2)}\ g_2^{(2)} \dots g_{2m}^{(2)}]$.

THE BOUND ON MINIMUM DISTANCE

The minimum distance d_m of a convolutional code determines the probability of decoding error when the code is being used with threshold decoding. A large value of d_m also reduces the amount of computation needed with sequential decoding. In this section we prove that the class of codes introduced in section 1 meets the Gilbert lower bound on d_m.

Theorem 1. There exists rate R = 1/2 codes of the type shown in figure 1 such that

$$\lim_{m\to\infty} \frac{d_m}{n_A} \geq H^{-1}(1 - R) = .11,$$

where $n_A = 2(m+1)$ is the code constraint length and $H(x) = -x \log x - (1 - x) \log (1 - x)$ is the binary entropy function.

Proof. An initial codeword $\underline{y}_m$ is defined as the first constraint length of transmitted digits for some codeword, i.e., $\underline{y}_m = [y_0^{(1)}\ y_0^{(2)}\ y_1^{(1)}\ y_1^{(2)} \dots y_m^{(1)}\ y_m^{(2)}]$. The proof follows a standard Gilbert bound argument (see, e.g., Massey [1]). We seek the smallest integer d_m such that

$$\left(\begin{matrix}\text{\# of initial codewords}\\ \text{with weight} \leq d_m\end{matrix}\right)\left(\begin{matrix}\text{\# of codes ruined by}\\ \text{each low weight codeword}\end{matrix}\right) \geq \left(\begin{matrix}\text{total \#}\\ \text{of codes}\end{matrix}\right), \qquad (2)$$

where a code is said to be ruined if it contains an initial codeword with weight $\leq d_m$. Clearly, there are

$$\sum_{j=0}^{d_m} \binom{n_A}{j} \leq 2^{n_A H\left(\frac{d_m}{n_A}\right)}$$

initial codewords of low weight. Also, since each of the generator sequences can be chosen arbitrarily, there are a total of 2^{4m+2} codes.

Now consider the number of codes ruined by each low weight initial codeword. For a given information sequence $\underline{x}_m = [x_0\ x_1 \cdots x_m]$ with $x_0 = 1$ and a given initial codeword $\underline{y}_m$, equations (1) imply that if we choose $g_{m+1}^{(1)}$, $g_{m+1}^{(2)}$, $g_{m+2}^{(1)}$, $g_{m+2}^{(2)}$, ..., $g_{2m}^{(1)}$, $g_{2m}^{(2)}$ arbitrarily, the remaining 2(m+1) g's are fixed. This can be seen by choosing the above mentioned g's arbitrarily and then noting that the equation for y_m fixes g_m, the equation for y_{m-1} fixes g_{m-1}, the equation for y_{m-2} fixes g_{m-2}, and so on. Hence a given information sequence and a given initial codeword ruins 2^{2m} codes. Since there are a total of 2^m information sequences with $x_0 = 1$, a given initial codeword can ruin at most 2^{3m} codes. (Some of the codes ruined by one information sequence may be the same as those ruined by another.) Therefore we seek the least integer d_m such that

$$2^{n_A H\left(\frac{d_m}{n_A}\right)} \times 2^{3m} \geq 2^{4m+2}. \tag{3}$$

This reduces to

$$(2m+2)H\left(\frac{d_m}{n_A}\right) \geq m+2, \tag{4}$$

and in the limit as $m \to \infty$

$$\lim_{m\to\infty} \frac{d_m}{n_A} \geq H^{-1}\left(\frac{1}{2}\right) = .11. \quad \text{(Q.E.D.)} \tag{5}$$

The generalization of this theorem to all rates can now be stated as a corollary.

Corollary 1. There exists simply instrumented time-varying convolutional codes with period $T = 2m+1$ such that

$$\lim_{m\to\infty} \frac{d_m}{n_A} \geq H^{-1}(1 - R).$$

THE BOUND ON FREE DISTANCE

The free distance d_f determines the error probability when the code is being used with either sequential or maximum likelihood decoding. In this section we prove a lower bound on d_f for the class of codes introduced in section 1 equal to twice the Gilbert bound for $R \geq 3/8$ and between one and two times the Gilbert bound for $R < 3/8$.

Theorem 2. There exists rate $R = 1/2$ codes of the type shown in figure 1 such that

$$\lim_{m\to\infty} \frac{d_f}{n_A} \geq 2H^{-1}(1 - R) = .22.$$

Proof. We follow a proof given by Costello [2] of a result originally due to Neumann [3], who obtained a lower bound on d_f for the class of non-systematic fixed (non-time-varying) codes. First consider only information sequences of length $h \leq m+2$ and divide each transmitted sequence $\underline{y}^{(1)}$ and $\underline{y}^{(2)}$ into two parts, part 1 consisting of the first $\frac{m+h}{2}$ digits and part 2 consisting of the last $\frac{m+h}{2}$ digits. (For convenience, we are assuming that $m+h$ is even.)

We will say that a code is ruined if it contains a codeword with weight d or less in either part 1 or part 2. Now let us rewrite the encoding equations for the case when $h = m+2$:

$$\begin{array}{l}
y_0 = x_0 g_0 \\
y_1 = x_1 g_2 + x_0 g_1 \\
\vdots \qquad\qquad\qquad\qquad \text{(part 1)} \\
y_{m-1} = x_{m-1} g_{2m-2} + x_{m-2} g_{2m-3} + \cdots + x_0 g_{m-1} \\
y_m = x_m g_{2m} + x_{m-1} g_{2m-1} + \cdots + x_1 g_{m+1} + x_0 g_m \\
\text{-------------------------------------} \\
y_{m+1} = x_{m+1} g_1 + x_m g_0 + \cdots + x_2 g_{m+3} + x_1 g_{m+2} \\
y_{m+2} = x_{m+1} g_2 + x_m g_1 + \cdots + x_2 g_{m+4} \\
\vdots \qquad\qquad\qquad\qquad \text{(part 2)} \\
y_{2m} = x_{m+1} g_m + x_m g_{m-1} \\
y_{2m+1} = x_{m+1} g_{m+1}
\end{array} \qquad (6)$$

Note that the encoding equations are symmetrical about the line dividing part 1 from part 2. For other values of $h < m+2$, the encoding equations exhibit the same symmetry. This symmetry implies that for a given information sequence of length h, the number of codes ruined by a codeword which has low weight in part 1 is the same as the number of codes ruined by a codeword with low weight in part 2. Hence we seek the smallest integer d such that

$$2\binom{\text{\#of codewords with}}{\text{weight} \leqq d \text{ in part 1}}\binom{\text{\#of codes ruined by each}}{\text{low weight codeword in part 1}} \geqq \binom{\text{total \#}}{\text{of codes}}. \qquad (7)$$

After eliminating all ruined codes, the remaining codes must have $d_f \geqq 2d$.

For a given information sequence $\underline{x}_{h-1} = [x_0 x_1 \cdots x_{m-1}]$ of length h with $x_0 = x_{h-1} = 1$, we can see from the encoding equations that a given codeword with low weight in part 1 (the first $\frac{m+h}{2}$ equations) fixes exactly $\frac{m+h}{2}$ positions of each generator sequence, i.e., $g_0^{(1)}, g_0^{(2)}, g_1^{(1)}, g_1^{(2)}, \ldots, g_{\frac{m+h}{2}-1}^{(1)}, g_{\frac{m+h}{2}-1}^{(2)}$ are fixed by the encoding equations. Hence a total of $4m+2-(m+h)$ positions can be chosen arbitrarily, and a given information sequence of length h and a given codeword with low weight in part 1 ruins 2^{3m+2-h} codes.

Since there are a total of 2^{h-2} information sequences of length h with $x_0 = x_{h-1} = 1$, a given codeword with low weight in part 1 can ruin at most 2^{3m} codes. Clearly there are

$$\sum_{j=0}^{d} \binom{m+h}{j} \leqq 2^{(m+h)H\left(\frac{d}{m+h}\right)}$$

codewords with low weight in part 1 and there are a total of 2^{4m+2} codes. Also, the total number of codes ruined by all information sequences of any length h, $1 \leq h \leq m+2$, is certainly less than or equal to the sum over h of the number ruined by all information sequences of given length h. Therefore we seek the least integer d such that

$$2\sum_{h=1}^{m+2} 2^{3m}2^{(m+h)H(\frac{d}{m+h})} \geq 2^{4m+2}. \tag{8}$$

We can simplify the left hand side of (8) by maximizing $(m+h)H(\frac{d}{m+h})$ over all h. Taking the derivative with respect to h, we obtain:

$$\frac{d}{dh}(m+h)H(\frac{d}{m+h}) = -\frac{\log(1-\frac{d}{m+h})}{\log 2} \geq 0 \tag{9}$$

if $0 \leq d \leq m+h$. (By definition d cannot be negative and a simple average distance upper bound implies that d cannot exceed m+h). Hence $(m+h)H(\frac{d}{m+h})$ is a non-decreasing function of h which is maximized by choosing $h = m+2$. Now replacing each term in the sum by its maximum value (this is equivalent to weakening the lower bound on d), we obtain:

$$2(m+2)2^{3m}2^{n_A H(\frac{d}{n_A})} \geq 2^{4m+2}. \tag{10}$$

This reduces to

$$\log 2(m+2) + n_A H(\frac{d}{n_A}) \geq m+2, \tag{11}$$

and in the limit as $m \rightarrow \infty$

$$\lim_{m\rightarrow\infty} \frac{d}{n_A} \geq H^{-1}(\frac{1}{2}) = .11. \tag{12}$$

Since $d_f \geq 2d$, we now have

$$\lim_{m\rightarrow\infty} \frac{d_f}{n_A} \geq .22. \tag{13}$$

Up to this point, information sequences have been restricted to lengths of m+2 or less. For information sequences of length $h > m+2$, each codeword consists of more than 2 constraint lengths of digits. But from theorem 1 we know that in the limit as $m \rightarrow \infty$ almost all codes have no codewords with weight less than the Gilbert bound over the first constraint length. This is also true for the last constraint length since the last constraint length of digits in a given codeword is an initial codeword in the reverse code, i.e., the code whose generator sequences are the reverse of the original code. (It is well known that all asymptotic bounds which use a Gilbert-like argument are true for almost all codes.) Hence for information sequences of length $h > m+2$, in the limit as $m \rightarrow \infty$ almost all codes have no codewords with weight less than twice the Gilbert bound ($= .22\, n_A$), and the theorem is proved. (Q.E.D.)

In the generalization of this theorem to all rates, maximizing the term in the sum corresponding to equation (8) becomes more difficult, and the bound is weakened for $R < 3/8$. This is not surprising since less improvement in free distance over minimum distance is expected for low rates than for high rates. The general result is now stated as a corollary.

Corollary 2. There exists simply instrumented time-varying convolutional codes with period $T = 2m+1$ such that

$$\lim_{m \to \infty} \frac{d_f}{n_A} \geqq \begin{cases} 2H^{-1}(1 - R) & \text{for } R \geqq 3/8 \\ \dfrac{2R(1 - 2^{2R-1})}{H(1 - 2^{2R-1}) + 2R - 1} & \text{for } R \leqq 3/8 \end{cases}$$

The bounds of corollaries 1 and 2 are sketched in figure 2.

DISCUSSION AND CONCLUSIONS

A simply instrumented class of time-varying codes with period $T = 2m+1$ has been constructed. These codes have been shown to meet reasonable lower bounds on both minimum distance and free distance, and hence constitute a good class of convolutional codes for use with either sequential or maximum likelihood decoding. It is still unknown whether these codes meet the strongest known bound on the entire class of time-varying codes (see Costello [2]). The difficulties in proving such a result appear to be roughly equivalent to those encountered in trying to prove the same result for the class of fixed codes. In fact it should be noted that the bounds obtained in this paper are also the best known bounds for fixed codes. (Note, however, that the class of codes presented here contains no fixed codes, and hence the results are not trivial.)

The fact that many of the best results obtained for convolutional codes require the assumption of time-varying generators has led to much speculation that there exists better time-varying codes than fixed codes for a given rate and constraint length. The complexities inherent in implementing time-varying encoders and in calculating their distance properties have precluded much research effort in this direction up until now. The simply instrumented class of codes introduced in this paper are certainly as good as fixed codes of the same rate and constraint length, and are probably better. Further research is presently being conducted on finding specific codes in this class, evaluating their distance properties, and comparing their performance with certain known fixed codes using sequential decoding.

REFERENCES

1. J. L. Massey, Threshold Decoding. Cambridge, Mass.: The M.I. T. Press, 1963.

2. D. J. Costello, "Free Distance Bounds for Convolutional Codes", IEEE

Trans. Inform. Th. (To appear).

3. B. Neumann, "Distance Properties of Convolutional Codes", M.S. Thesis, E.E. Dept., M.I.T., 1968.

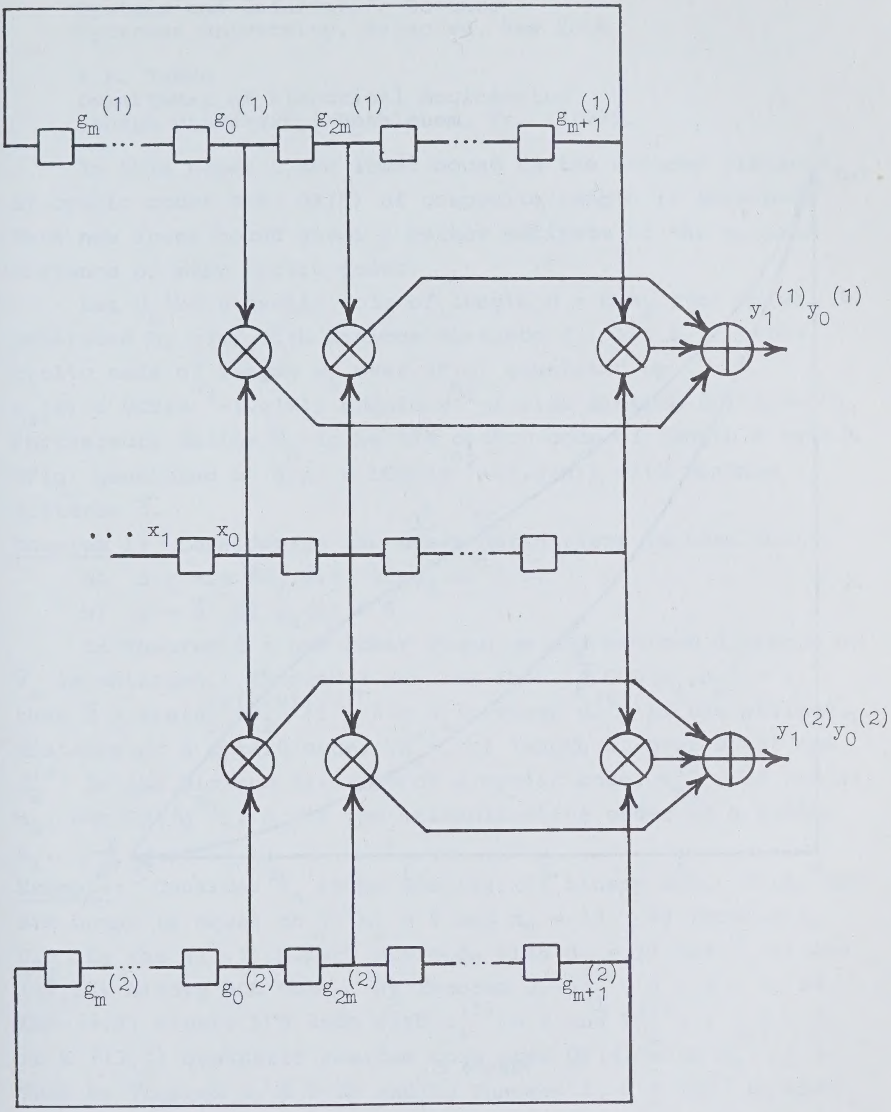

$g_m^{(1)}$
$g_0^{(1)}$
$g_{2m}^{(1)}$
$g_{m+1}^{(1)}$
$y_1^{(1)} y_0^{(1)}$
$\ldots x_1$
x_0
$y_1^{(2)} y_0^{(2)}$
$g_m^{(2)}$
$g_0^{(2)}$
$g_{2m}^{(2)}$
$g_{m+1}^{(2)}$

Figure 1

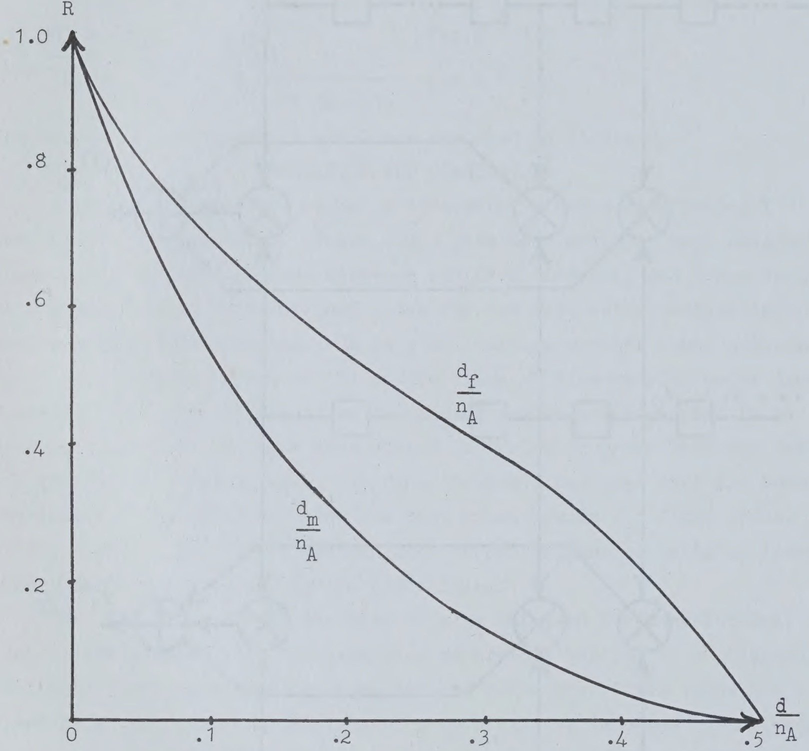


Figure 2

ON THE MINIMUM DISTANCE STRUCTURE OF CYCLIC CODES OF COMPOSITE LENGTH

C.R.P. HARTMANN
Systems and Information Science
Syracuse University, Syracuse, New York

K.K. TZENG
Department of Electrical Engineering
Lehigh University, Bethlehem, Pa. 18015.

In this paper a new lower bound on the minimum distance of cyclic codes over GF(q) of composite length is obtained. This new lower bound gives a better estimate of the minimum distance of many cyclic codes.

Let V_n be a cyclic code of length $n = n_1 n_2$ over GF(q) generated by g(x) with minimum distance d. Let V_{n_2} be the cyclic code of length n_2 over GF(q) generated by $g_2(x) \equiv \mathrm{GCD}(x^{n_2}-1, g(x))$ modulo $x^{n_2}-1$ with minimum distance d_2. Furthermore define $\bar{V}_n$ to be the cyclic code of length n over GF(q) generated by $\bar{g}(x) = \mathrm{LCM}((x^{n_2}-1), g(x))$ with minimum distance $\bar{d}$.

Theorem 1: Considering the above definitions we have that

a) $d \geq \min(d_2, \bar{d})$* if $g_2(x) \not\equiv 0$.

b) $d = \bar{d}$ if $g_2(x) \equiv 0$.

In Theorem 2 a new lower bound on the minimum distance of $\bar{V}_n$ is obtained. Theorem 2 assures that if $\mathrm{GCD}(n_1, n_2) = 1$, then $\bar{d} \geq \min\{d_1^{(\theta)} d_2^{(\theta)} \mid 1 \leq \theta < n_1\}$, where $d_1^{(\theta)}$ is the minimum distance of a cyclic code, $V_{n_1}^{(\theta)}$, of length n_1 over GF(q) and $d_2^{(\theta)}$ is the minimum distance of a cyclic code, $V_{n_2}^{(\theta)}$, of length n_2, over $GF(q^{m_1})$, m_1 is the multiplicative order of q modulo n_1.

Example: Consider V_n to be the (39,15) binary BCH. Thus, the BCH bound is equal to 7, $n_1 = 3$ and $n_2 = 13$. By Theorem 1, V_{n_2} is the (13,1) binary BCH code with $d_2 = 13$ and $\bar{V}_n$ is the (39,14) binary BCH code. By Theorem 2, $V_{n_1}^{(\theta)}$, $1 \leq \theta < 3$, is the (3,2) binary BCH code with $d_1^{(\theta)} = 2$ and $V_{n_2}^{(\theta)}$, $1 \leq \theta < 3$, is a (13,7) quadratic residue code over GF(4) with $d_2^{(\theta)} \geq 5$. Thus by Theorem 2, $\bar{d} \geq 10$ and by Theorem 1, $d \geq 10$. In this case, it is known that d = 10. Readers interested in more details should refer to the full paper [1].

---

*min(a,b) denotes the minimum of a and b.

[1] C.R.P. Hartmann and K.K. Tzeng, "On Some Classes of Cyclic Codes of Composite Length," submitted to publication in the IEEE Trans. on Information Theory.

A NOTE ON THE MINIMUM DISTANCE STRUCTURE OF CYCLIC CODES

C.R.P. HARTMANN
Systems and Information Science
Syracuse University, Syracuse, New York

In this note a modified version of the Mattson-Solomon formulation [1] is applied to cyclic codes and better estimates of the minimum distance, d, of many cyclic codes can be obtained.

Let $v(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1}$ be a code vector of a cyclic code, V_n, of length n generated by $g(x) \in GF(q)[x]$. Let $x^n - 1 = g(x)f(x)$ and let us define $E(\beta) = \{e_1, e_2, \ldots, e_k \mid f(\beta^{e_i}) = 0 \quad 1 \le i \le k, \text{ and } e_j < e_{j+1}, 1 \le j \le k-1\}$, where β is a primitive n^{th} root of unity. Mattson and Solomon [1] showed that for a given $v(x)$ there exists a polynomial $g_v(x)$ such that $a_i = g_v(\beta^i)$, where $g_v(x) = c_1x^{e_1} + c_2x^{e_2} + \ldots + c_kx^{e_k}$, $e_i \in E(\beta)$, $c_i \in GF(q^m)$ and if $e_i \equiv qe_j$ modulo n then $c_i = c_j^{\,q}$. Let us now define $\Delta(i,i+1) = e_{i+1} - e_i$, $1 \le i \le k-1$ and $\Delta(k,k+1) = n + e_1 - e_k$. If $g_v(x)$ has n-w roots among the n^{th} roots of the unity then wght $(v(x)) = w$*. The BCH bound, d_0, can be obtained as follows: let $s, 1 \le s \le k$, be such that $d_0 = \Delta(s,s+1) \ge \Delta(i,i+1)$, $1 \le i \le k$, and consider the polynomial $p_v(x) \equiv x^{n-e_{s+1}} g_v(x)$ modulo $x^n - 1$ so wght $(v(x)) \ge n - \deg(p_v(x)) = d_0$**. Since this is valid for all $v(x)$, then $d \ge d_0$.

The basic idea of this paper is to define, for each $v(x)$, a polynomial $m_v(x)$ with the following properties:

(1) if $g_v(\beta^i) = 0$, then $m_v(\beta^i) = 0$.

(2) $\deg(p_v(x)) > \deg(m_v(x))$.

Since $g_v(x)$ cannot have more roots among the n^{th} roots of unity than $m_v(x)$, then we can conclude that if $m_v(x) \not\equiv 0$, then wght $(v(x)) \ge n - \deg(m_v(x))$.

Example: Consider the quadratic residue (11,6) code over GF(4). For this code $E(\beta) = \{0,2,6,7,8,10\}$ and $d_0 = 4$. Let $v(x) \not\equiv 0$ be a code vector, thus

$g_v(x) = c_1 + c_2x^2 + c_3x^6 + c_4x^7 + c_5x^8 + c_6x^{10}$.

---

*wght (b(x)) denotes the weight of b(x).

**deg(b(x)) denotes the degree of b(x).

So, $g_v^2(x) = c_1^2+c_3^2x+c_4^2x^3+c_2^2x^4+c_5^2x^5+c_6^2x^9$. Define $m_v(x)$ as: $m_v(x) \equiv (c_5^2x^5g_v(x)+c_2x^2g_v^2(x))$ modulo $x^{11}+1$. So,

$$m_v(x) = (c_5^2c_3+c_2c_6^2) + c_5^2c_4x + (c_5^3+c_2c_4^2)x^2 +$$
$$+ c_2c_3^2x^3+c_5^2c_6x^4 + (c_5^2c_1+c_2c_4^2)x^5+c_2^3x^6.$$

If $c_5 = 0$ or $c_2 = 0$ then $c_2=c_3=c_4=c_5=c_6 = 0$ and wght $(v(x)) = 11$. If $m_v(x) \equiv 0$ then $c_2 = 0$ and wght $(v(x)) = 11$. If $m_v(x) \not\equiv 0$ then $g_v(x)$ has at most 6 roots among the 11th roots of unity and wght $(v(x)) \geq 5$. Since this is valid for any $v(x)$, then $d \geq 5$. In this case, it is known that $d = 5$.

The cyclic codes with multiple sets of consecutive roots of the generator polynomial are very suitable to the application of this technique.

[1] H.F. Mattson and G. Solomon, "A New Treatment of Bose-Chaudhuri Codes," SIAM, Vol. 9, pp. 654-699, December, 1961.

SOME NOTES ON ITERATIVE DECODING

K. K. TZENG
Dept. of Electrical Engineering
Lehigh University
Bethlehem, Pennsylvania 18015

C. R. P. HARTMANN
Systems and Information Science
Syracuse University
Syracuse, New York 13210

and

R. T. CHIEN
Coordinated Science Laboratory
University of Illinois
Urbana, Illinois 61801

ABSTRACT

Two aspects pertaining to the Berlekamp Iterative Algorithm are investigated. Firstly, it is shown that in decoding a t-error-correcting BCH code, a detection criterion exists so that the iterative decoder can stop the iteration at the completion of the n-th step where $n = \mathrm{MIN}[2t, t+n'-D(n')]$. Furthermore, when there are e errors, $e < t$, then $n'-D(n') = e$ or $e-1$ and thus the iteration may be terminated in such cases at latest at the completion of the (t+e)th step. Secondly, it is proved that the iterative decoding of the syndrome sequence in the reversed order is not only meaningful but is also very useful in the decoding of a class of binary reversible codes beyond their BCH bound.

I. INTRODUCTION

The problem of decoding a t-error-correcting BCH code is generally considered as one of finding the e error locations X_i and the e error values Y_i for $i = 1, 2, \ldots, e$ from a given set of 2t syndrome terms $S_1, S_2, \ldots, S_{2t}$ where $S_j = \sum_{i=1}^{e} Y_i X_i^j$ [1], [2]. Berlekamp [2] has shown that this is equivalent to the solution of the key equation

$$(1 + S(z))\sigma(z) \equiv \omega(z) \quad \text{modulo } z^{2t+1}$$

where

$$S(z) = \sum_{k=1}^{\infty} S_k z^k \text{ and } \sigma(z) = \prod_{i=1}^{e} (1-X_i z) = 1 + \sigma_1 z + \sigma_2 z^2 + \cdots + \sigma_e z^e$$

and has also introduced a procedure, now well known as the Berlekamp Iterative Algorithm, to find $\sigma(z)$ and $\omega(z)$ iteratively.

In this paper, we are concerned in the first place with the number of sufficient steps of iteration toward finding $\sigma(z)$ and $\omega(z)$. Given 2t syndrome terms $S_1, S_2, \ldots, S_{2t}$, the iterative decoder in general has been required to go through 2t steps of iteration to find $\sigma(z)$ and $\omega(z)$. When there are t or more errors, then it appears necessary to have 2t steps of iteration. However, if the number of errors is $e < t$, then 2e steps of iteration is all that is necessary since the decoder will have found $\sigma(z)$

This work is supported by the National Science Foundation under Grants GK 24879 and GK 5647; auxiliary support is provided by the Joint Services Electronics Program under Contract DAAB-07-67-c-0199.

and $\omega(z)$ already in 2e steps of iteration. Furthermore, 2e steps will also be sufficient in such cases but the decoder will not stop the iterative computation unless some means for detecting this situation are provided.

In the following section, we show that detection criterion does exist so that an earlier termination of the iterative computation is possible. Furthermore, if there are e errors, $e < t$, then it is shown that the decoder can stop the iteration at latest at the completion of the $(t+e)$th step of iteration. Since some computation which is more complicated in the later stages of iteration is eliminated with such detection, considerable decoding time will then be saved.

Secondly, in applying this algorithm, it has usually been assumed that the sequence of syndrome terms is necessarily to be used in the order of increasing subscript. In this paper, we show in Section III, however, that the iterative decoding of the syndrome sequence in the reversed order is not only meaningful but also very useful as is illustrated in the decoding of a class of binary reversible codes beyond their BCH bound.

For the details of the iterative algorithm, the reader is referred to [2-5].

II. THE DETECTION OF SUFFICIENT ITERATION

The detection criterion for sufficient iteration is now formally stated in the following theorem.

Theorem 1. Given 2t syndrome terms $S_1, S_2, \ldots, S_{2t}$, then the decoder can stop at the completion of the nth step of iteration whenever for $n < 2t$, $\Delta(n) = 0$ and $n = t + n' - D(n')$. Furthermore, if there are $e \leqslant t$ errors in the received vector, then $n' - D(n') = e$ or $e - 1$, therefore, the iteration can be stopped at latest at the $(t + e)$th step.

Proof.

It is noted that $D(n) = \mathrm{Max}[\deg \sigma^{(n)}, \deg \omega^{(n)}]$ is a monotonically increasing function and that $D(k+1) = \mathrm{Max}[D(k), k-n' + D(n')]$ whenever $\Delta(k) \neq 0$. Suppose for $n < 2t$, we have $\Delta(n) = 0$ and $n = t + n' - D(n')$. Then either $\Delta(k) = 0$ for all $n > k$ which means the correct solution has been obtained or else there exists $\Delta(k) \neq 0$ for $k > n$. In the later case, for the first k such that $\Delta(k) \neq 0$, we have $D(k+1) \geqslant k-n' + D(n') > t$ indicating more than t errors in the received vector. Therefore, whenever $\Delta(n) = 0$ and $n = t + n' - D(n')$ for $n < 2t$, the decoder can stop before the 2t-th step of iteration with the confidence that either the correct solution has already been obtained or else that there are more than t errors in the received vector.

Next, we show that when there are $e \leqslant t$ errors, the value of $n' - D(n')$ after the correct answers of $\sigma(z)$ and $\omega(z)$ are obtained at the 2eth step is equal to e or e - 1.

First of all, we have

$$D(2e) = \mathrm{Max}[D(2e-1), (2e-1) - n'' + D(n'')] = e$$

(1) If $D(2e-1) = 2e-1 - n'' + D(n'') = e$, then

$$2e-1 - D(2e-1) = n'' - D(n'') = e-1, \text{ then}$$

$$n' = n'' \text{ and } n' - D(n') = e - 1.$$

(2) If $D(2e-1) < e$ and $2e-1 - n'' + D(n'') = e$, then

$$2e-1 - D(2e-1) > e - 1 \text{ and } n'' - D(n'') = e - 1.$$

Then $n' = 2e-1$ if in addition $\Delta(2e-1) \neq 0$.

Now $D(2e-1) = \text{Max}\,[D(2e-2),\ 2e-2 - n''' + D(n''')] < e$.

If $n'' = 2e-2$, then $D(2e-2) = e-1$. If $n'' = n'''$, then $2e - 2 - n''' + D(n''') = e - 1$. Hence $D(2e-1) = e - 1$ and we have $n' - D(n') = e$.

(3) If $D(2e-1) = e$ and $2e - 1 - n'' + D(n'') < e$, then $2e - 1 - D(2e-1) = e - 1$ and $n'' - D(n'') > e - 1$. Then $n' = n'' < 2e - 1$, $n' - D(n') > e - 1$. Hence $D(n') < e$.

Now consider

$$D(2e-1) = \text{Max}[D(2e-2),\ 2e-2-n''' + D(n''')] = e.$$

Following (1), we have $n' - D(n') = e - 2$ which is a contradiction. Following (2), we have $n' - D(n') = e - 1$ which is also a contradiction. Following (3), we have $D(2e-2) = e$. Then recursively it can be shown that $D(2e-3) = e \cdots$, and finally $D(n') = e$ which is again a contradiction. Therefore, we have $n' - D(n') = e$ or $e - 1$. Then $n = t + n' - D(n') = t + e$ or $t + e-1$. Hence the iterative computation can be stopped at latest at the $(t+e)$th step when the number of errors is $e \leqslant t$.

Q.E.D.

III. ITERATIVE DECODING OF THE REVERSE SEQUENCE

First of all, we claim that when the iterative decoding procedure is applied to the reverse sequence $S_{2t}, S_{2t-1}, \cdots, S_1$, then the reciprocal polynomial of $\sigma(z)$ will be obtained.

That this should be so is simply traced from the invariance of the generalized Newton's Identities under the substitutions

$$S_s \to S_{2t+1-s} \qquad \text{for } s = t+1, t+2, \cdots, 2t$$

and

$$\sigma_i \to \frac{\sigma_{t-i}}{\sigma_t} \qquad \text{for } i = 1, 2, \cdots, t$$

since the generalized Newton's Identities are

$$S_s + S_{s-1}\sigma_1 + \cdots + S_{s-t+1}\sigma_{t-1} + S_{s-t}\sigma_t = 0$$

$$\text{for } s = t+1, t+2, \cdots, 2t, \cdots$$

which are unaltered under the above substitutions.

With these substitutions, the key equation becomes

$$(1 + S_{2t}z + S_{2t-1}z^2 + \cdots + S_2 z^{2t+1} + S_1 z^{2t} + \cdots)(1 + \frac{\sigma_{t-1}}{\sigma_t}\ z$$

$$+ \frac{\sigma_{t-2}}{\sigma_t} z^2 + \cdots + \frac{\sigma_1}{\sigma_t} z^{t-1} + \frac{1}{\sigma_t} z^t)$$

$$= 1 + (S_{2t} + \frac{\sigma_{t-1}}{\sigma_t})z + \cdots + (S_{t+1} + S_{t+2} \frac{\sigma_{t-1}}{\sigma_t} + \cdots +$$

$$S_{2t} \frac{\sigma_1}{\sigma_t} + \frac{1}{\sigma_t}) z^t \qquad \text{mod } z^{2t+1}$$

Let

$$\frac{\tilde{\sigma}(z)}{\sigma_t} = 1 + \frac{\sigma_{t-1}}{\sigma_t} z + \frac{\sigma_{t-2}}{\sigma_t} z^2 + \cdots + \frac{\sigma_1}{\sigma_t} z^{t-1} + \frac{1}{\sigma_t} z^t$$

and

$$\frac{\bar{\omega}(z)}{\sigma_t} = 1 + (S_{2t} + \frac{\sigma_{t-1}}{\sigma_t}) z + \cdots + (S_{t+1} + S_{t+2} \frac{\sigma_{t-1}}{\sigma_t}$$

$$+ \cdots + S_{2t} \frac{\sigma_1}{\sigma_t} + \frac{1}{\sigma_t}) z^t$$

Then

$$\tilde{\sigma}(z) = z^t + \sigma_1 z^{t-1} + \cdots + \sigma_{t-1} z + \sigma_t$$

$$= \prod_{i=1}^{t} (z - X_i)$$

which is the reciprocal polynomial of $\sigma(z)$.

With some manipulations, it can also be shown that

$$\bar{\omega}(X_i) = - Y_i X_i^{2t} \prod_{j \neq 1} (X_i - X_j)$$

Hence

$$Y_i = \frac{- \bar{\omega}(X_i)}{X_i^{2t} \prod_{j \neq i} (X_i - X_j)}$$

Therefore, $\bar{\omega}(z)$ is the error evaluation polynomial in the case of reverse sequence iterative decoding.

These findings can be summarized in the following theorem.

Theorem 2: Let $S_1, S_2, \cdots, S_{2t}$ be a given syndrome sequence. If the iterative algorithm is applied to the reversed sequence $S_{2t}, S_{2t-1}, \cdots, S_1$, then at the termination of the algorithm we have $\sigma^{(2t)} = \frac{\tilde{\sigma}(z)}{\sigma_t}$ and $\omega^{(2t)} = \frac{\bar{\omega}(z)}{\sigma_t}$. The error locations are roots of $\tilde{\sigma}(z)$ and the error magnitudes can be found from the given error evaluation through $\bar{\omega}(z)$.

One of the interesting applications of reverse sequence iterative decoding is in the decoding of a class of binary reversible cyclic codes beyond their BCH bound.

Let g(z), not divisible by 1 + z, be a self-reciprocal polynomial of least degree with coefficients in GF(2) such that for some nonzero element β in GF(2^m), $\beta^{\pm 1}$, $\beta^{\pm 2}$, $\cdots$, $\beta^{\pm(2t-1)}$, $\beta^{\pm 2t}$ are among the roots of g(z). It has been shown by Tzeng and Hartmann [6], that the minimum distance of the reversible code generated by g(z) is equal to the BCH bound $d_o = 2t+1$ when d_o divides the length n of the code. If d_o does not divide n, then the minimum distance is at least $d_o + 2$.

To decode this class of reversible codes, it is noted that two sequences of successive syndrome terms are available. Let us call $S_1, S_2, \cdots, S_{2t}$ the forward sequence and $S_{-1}, S_{-2}, \cdots, S_{-2t}$ the reverse sequence. It should be noted that for the codes under consideration, we have for both sequences $S_{2k} = S_k^2$. Hence in the iterative decoding of both sequences, we have $\Delta(n) = 0$ for n odd. For decoding up to t errors, either sequence can be used in the iterative decoding scheme. However, when t + 1 errors are to be corrected, then both sequences must be used.

Let us assume that there are t + 1 errors in the received vector. Since S_{2t+1} is not known, the polynomial $\sigma^{(2t+1)}$ calculated by the forward sequence decoding algorithm will contain S_{2t+1} as an unknown. The problem of finding $\sigma^{(2t+2)} = \sigma^{(2t+1)}$ is then equivalent to the problem of solving for S_{2t+1}. Similarly, for the reverse sequence, we have $\dfrac{\tilde{\sigma}^{(2t+1)}}{\sigma_{t+1}^{(2t+1)}}$ containing the unknown $S_{-(2t+1)}$. Now

$$\sigma(z) = \sigma^{(2t+2)} = \sigma^{(2t+1)} = 1 + \sigma_1 z + \sigma_2 z^2 + \cdots + \sigma_{t+1} z^{t+1}$$

$$\frac{\tilde{\sigma}(z)}{\sigma_{t+1}} = \frac{\tilde{\sigma}^{(2t+2)}}{\sigma_{t+1}^{(2t+2)}} = \frac{\tilde{\sigma}^{(2t+1)}}{\sigma_{t+1}^{(2t+1)}} = 1 + \frac{\sigma_t}{\sigma_{t+1}} z + \frac{\sigma_{t-1}}{\sigma_{t+1}} z$$

$$+ \cdots + \frac{1}{\sigma_{t+1}} z^{t+1} .$$

Then S_{2t+1} and also $S_{-(2t+1)}$ can be solved from the relationships existing between the coefficients of these two polynomials.

We remark that when the minimum distance of the code is at least 2t+3 then the solutions for S_{2t+1} and $S_{-(2t+1)}$ will be unique. If the minimum distance is just 2t+1, then S_{2t+1} and $S_{-(2t+1)}$ may have more than one solution.

Example: Let α be a primitive element of GF(2^5) where $\alpha^5 + \alpha^2 + 1 = 0$ and consider the (31,11) reversible binary cyclic code that has $\alpha^{\pm 1}$, $\alpha^{\pm 2}$, $\alpha^{\pm 3}$ and $\alpha^{\pm 4}$ among the roots of its generator polynomial.

Suppose the syndrome terms calculated at the decoder for this code are

$$S_1 = \alpha^{18},\ S_2 = S_1^2 = \alpha^5,\ S_3 = \alpha^3,\ S_4 = S_2^2 = \alpha^{10}$$

$$S_{-1} = \alpha^5,\ S_{-2} = S_{-1} = \alpha^{10},\ S_{-3} = \alpha^{18},\ S_{-4} = S_{-2} = \alpha^{20}$$

Using the iterative algorithm [2-5] for the forward sequence we get

$$\begin{aligned} \sigma^{(6)} &= \sigma^{(5)} = \sigma^{(4)} - \Delta(4)\Delta(2)^{-1}z^{4-2}\sigma^{(2)} \\ &= 1 + \alpha^{18}z + (\alpha^{28} + \alpha^{20}S_5)z^2 + (\alpha^{21} + \alpha^7 S_5)z^3 . \end{aligned}$$

For the reverse sequence, we have

$$\begin{aligned} \sigma^{(6)} &= \sigma^{(5)} = \sigma^{(4)} - \Delta(4)\Delta(2)^{-1}z^{4-2}\sigma^{(2)} \\ &= 1 + \alpha^5 z + (\alpha^{15} + \alpha^{18}S_{-5})z^2 + (\alpha^4 + \alpha^{23}S_{-5})z^3 \end{aligned}$$

Thus

$$\sigma(z) = 1 + \alpha^{18}z + (\alpha^{28} + \alpha^{20}S_5)z^2 + (\alpha^{21} + \alpha^7 S_5)z^3$$

$$\frac{\tilde{\sigma}(z)}{\sigma_3} = 1 + \alpha^5 z + (\alpha^{15} + \alpha^{18}S_{-5})z^2 + (\alpha^4 + \alpha^{23}S_{-5})z^3$$

From which we have

$$S_5 = \alpha^{30}$$

Then

$$\begin{aligned} \sigma(z) &= 1 + \alpha^{18}z + \alpha^4 z^2 + \alpha^{30} z^3 \\ &= (1 + z)\,(1 + \alpha^9 z)\,(1 + \alpha^{21} z) \end{aligned}$$

Therefore, three errors occurred at the first, 10th and 22nd locations.

VI. SUMMARY AND CONCLUSIONS

In this paper, we have shown that the iterative decoder can stop its iteration when the condition $\Delta(n) = 0$ and $n = t + n' - D(n')$ is detected. We have also shown that $n' - D(n') = e$ or $e - 1$. Hence the iteration can be stopped at latest at the (t+e)-th step. Since such detection can eliminate some complicated computation at the later stages of iteration, considerable decoding time can thus be saved. On decoding the syndrome sequence in the reversed order, we have shown that this is not only meaningful but is also a very natural technique to be coupled with the forward sequence iterative decoding to decode a class of binary reversible codes beyond their BCH bound.

REFERENCES

[1] W. W. Peterson, Error Correcting Codes, M.I.T. Press, and Wiley, New York; 1961.

[2] E. R. Berlekamp, Algebraic Coding Theory, McGraw-Hill, New York; 1968.

[3] J. L. Massey, "Shift Register Synthesis and BCH Decoding", IEEE Transactions on Information Theory, Vol. IT 15, pp. 122-127, January, 1969.

[4] R. T. Chien and K. K. Tzeng, "On Iterative Decoding of BCH Codes", Proceedings of the Eleventh Midwest Symposium on Circuit Theory, University of Notre Dame, Notre Dame, Indiana, May, 1968, pp. 599-600.

[5] K. K. Tzeng, "On Iterative Decoding of BCH Codes and Decoding beyond the BCH Bound", Report R-404, Coordinated Science Laboratory, University of Illinois, Urbana; January, 1969.

[6] K. K. Tzeng and C. R. P. Hartmann, "On the Minimum Distance of Certain Reversible Cyclic Codes", IEEE Transactions on Information Theory, Vol. IT-16, pp. 644-646, September, 1970.

DETERMINING CODE CAPABILITIES FOR NONRANDOM ERRORS

DOUGLAS A. ANDERSON
Bell Telephone Laboratories
Naperville, Illinois

GERNOT METZE
University of Illinois
Urbana, Illinois

Abstract: A 2-component "polarized distance" is introduced and is shown to be useful for determining the error-detecting/correcting capabilities of binary block codes in cases where it is important to consider the number of 0-errors and the number of 1-errors separately. The detection and correction of asymmetric and unidirectional errors are considered in detail and some specific codes are analyzed using this polarized distance technique.

I. INTRODUCTION

When the error probabilities of a particular environment are not random but have some directional or positional dependencies, then the Hamming distance is insufficient to measure code capabilities. Marcus [11] showed that different codes with the same minimum Hamming distance could have a much different probability of undetected error in an asymmetric channel and suggested a 2-component "polarized distance" for this purpose. Freiman [3] formalized this idea with a similar 2-component "vector distance." Using this notation in [3], he described the necessary and sufficient conditions for a specific type of asymmetric error detection/correction which rectified mistakes made in an earlier paper [10]. The polarized distance defined in this paper is the same as Freiman's vector distance and his results are extended to the detection and correction of errors with arbitrary directional dependencies, that is, to any case in which it is important to consider the number of 0-errors and the number of 1-errors separately. In this work, a 0-error is defined to be an error that causes a 1 to be changed to a 0; similarly, a 1-error causes a 0 to be changed to a 1.

II. POLARIZED WEIGHT AND DISTANCE

The mathematical model used to formalize this distance concept is the set of n-dimensional vectors with integer components. Addition is component-wise integer addition (not modulo 2). The set of 2^n vectors in the model that have only 0s and 1s for components forms the (physically) realizable space. The set of 3^n vectors that have only 0s, 1s, and -1s for components forms the set of error representations. For example, if

$$r = 10010010 \quad \text{is the result received}$$

and $$x = 11011000 \quad \text{was the code word sent}$$

then $$e = r - x = 0\bar{1}00\bar{1}010 \quad \text{is the error pattern}$$

that was added to x ($\bar{1}$ represents -1).

Definition: The polarized weight, $W_P(e)$, of an error e is $W_P(e) = (w^+, w^-)$, where $w^+(e)$ is the number of positive components of e and $w^-(e)$ is the number of negative components of e. From the previous example, $e = 0\bar{1}00\bar{1}010$ and $W_P(e) = (1, 2)$.

Definition: The polarized distance, $D_P(x, y)$, between two code words x and y is $D_P(x, y) = W_P(x - y)$. Again from the previous example, $D_P(r, x) = W_P(r - x) = W_P(e) = (1, 2)$.

To clarify the discussion of the polarized distance concept it will be useful to define the function $d_{10}(x, y)$ as the number of corresponding bit positions in which x contains a 1 and y contains a 0. The functions $d_{01}(x, y)$, $d_{00}(x, y)$, $d_{101}(x, r, y)$, and so on will be defined similarly. From these definitions it follows that:

$$d_{10}(x, y) = w^+(x - y)$$

and

$$d_{01}(x, y) = w^-(x - y)$$

and

$$D_P(x, y) = (d_{10}, d_{01}) .$$

As suggested by the above equations, the two functions d_{10} and d_{01} will be of considerable importance in developing the ideas of polarized distance. From their definitions, it is obvious that:

$$d_{10}(x, y) \geq 0 \qquad (1)$$

and

$$d_{01}(x, y) \geq 0 \qquad (2)$$

with simultaneous equality if and only if x = y, and

$$d_{10}(x, y) = d_{01}(y, x) , \qquad (3)$$

the antisymmetric condition. Also note that:

$$d_{10}(x, y) = d_{100}(x, r, y) + d_{110}(x, r, y)$$

$$d_{10}(x, r) = d_{100}(x, r, y) + d_{101}(x, r, y)$$

and

$$d_{10}(r, y) = d_{010}(x, r, y) + d_{110}(x, r, y) .$$

Therefore, it follows that

$$d_{10}(x, r) + d_{10}(r, y) = d_{10}(x, y) + d_{101}(x, r, y) + d_{010}(x, r, y) .$$

Similarly it can be shown that

$$d_{01}(x, r) + d_{01}(r, y) = d_{01}(x, y) + d_{101}(x, r, y) + d_{010}(x, r, y) .$$

Let δ be defined as

$$\delta = d_{101}(x, r, y) + d_{010}(x, r, y) . \qquad (4)$$

Then the above equations can be written as

$$d_{10}(x, y) = d_{10}(x, r) + d_{10}(r, y) - \delta \qquad (5)$$

and
$$d_{01}(x, y) = d_{01}(x, r) + d_{01}(r, y) - \delta . \qquad (6)$$

Since $\delta \geq 0$, equations (5) and (6) imply two important triangle inequalities:

$$d_{10}(x, y) \leq d_{10}(x, r) + d_{10}(r, y)$$

and
$$d_{01}(x, y) \leq d_{01}(x, r) + d_{01}(r, y) .$$

Although, as seen above, d_{10} and d_{01} do not satisfy the three conditions of a metric, they do satisfy three very similar conditions which allow them, when considered as a pair (d_{10}, d_{01}), to take on the aspects of a distance.

III. POLARIZED DISTANCE SPACE

To study the capabilities of codes using the polarized distance concept it will be essential to display this pair of values in a 2-dimensional graphical presentation, as shown in Figure 1, which displays the possible polarized distance values $D_P = (d_{10}, d_{01})$ in a code of length n. Note that

$$d_{10} + d_{01} + d_{00} + d_{11} = n .$$

Thus, the sum of $d_{10} + d_{01}$ must satisfy the following inequality:

$$0 \leq d_{10} + d_{01} \leq n$$

which explains the triangular shape of Figure 1.

The polarized distance space, illustrated in Figure 1, is also a 2-dimensional display of all possible polarized weight classes. For example the point (2, 1) represents all error patterns of polarized weight (2, 1), that is, errors with two 1s and a single $\overline{1}$. There are $(n + 2)(n + 1)/2$ points in the polarized distance space, and thus, the polarized weight function partitions the set of 3^n error patterns into this many classes. A region of the polarized distance space will be defined as a set of such points. A region thus contains a certain number of weight classes and can be thought of as representing the error patterns in those weight classes. It will be necessary to identify two types of regions: one region called a correction-region or C-region will be identified with solid dots, and the other region called a detection-region or D-region identified with open dots. An example is shown in Figure 2.

The shadow of a point (a_1, a_2) will be defined to be the set of points $(a_1 - 1, a_2 - 1)$, $(a_1 - 2, a_2 - 2)$, and so on, the last point being $(a_1 - a_2, 0)$ if $a_1 \geq a_2$ or $(0, a_2 - a_1)$ if $a_1 \leq a_2$. The shadow of a point forms a diagonal line of points in the polarized distance space "behind" the given point when viewed from the point (∞, ∞). A shaded region will be defined to be one that contains the shadow of each of its elements. Thus an internal point of a shaded region lies in the shadow of some boundary point of the region. Two points are said to be adjacent if they lie vertically or horizontally next to each other in the polarized distance space. A disconnected region is defined to be one that contains at least two points such that it is impossible to traverse from one point to the other over adjacent points in the region. For example the point (1, 1) and its shadow (0, 0) form a disconnected region. If a region is not disconnected, then it is said to be connected.

To simplify matters, the following work will consider only C and D regions such that their union is a shaded connected region contains the point (0, 0). Such region are the only ones of importance since we are ultimately concerned with the correction and detection of errors with low polarized weight. In this light, the point (0, 0) which represents no error is always a member of the C region and will be denoted by x, which may be thought of as the polarized distance from code word x to itself, that is, $D_P(x, x) = (0, 0)$. Now consider another code word y at point $D_P(x, y)$ which will be denoted by y for short. Next consider point y as the origin of another polarized distance space with axes labeled d'_{10} and d'_{01}, except let the increasing d'_{10}-axis point in the opposite direction as the increasing d_{01}-axis and, similarly, let the increasing d'_{01}-axis point in the opposite direction as the increasing d_{10}-axis. The original C and D regions now have antisymmetric counterparts in this new space which will be called C′ and D′. An example is shown in Figure 2. Note that from equation (3),

$$d_{10}(x, y) = d'_{01}(y, x) \tag{7}$$

and

$$d_{01}(x, y) = d'_{01}(y, x) \; . \tag{8}$$

Using the above relations in equations (5) and (6) yields

$$d_{10}(x, y) = d_{10}(x, r) + d'_{01}(y, r) - \delta \tag{9}$$

and

$$d_{01}(x, y) = d_{01}(x, r) + d'_{10}(y, r) - \delta \; . \tag{10}$$

The above two equations will be very important in the development of the polarized distance concept.

IV. A POLARIZED METRIC

It is instructive to define a primed counterpart weight and distance as $W'_P(e) = (w^-, w^+)$ and $D'_P(y, x) = W'_P(y - x)$.

Since $w^{+}(x - y) = w^{-}(y - x)$, it can be seen from the above definitions that $D'_P(y, x) = (d'_{01}, d'_{10})$. Remembering the definition $D_P(x, y) = (d_{10}, d_{01})$, it follows that the polarized distance satisfies the following three conditions. Equations (1) and (2) may be written as

1) $$D_P(x, y) \geq (0, 0)$$

with equality if and only if $x = y$, where $(a_1, a_2) \geq (0, 0)$ means that $a_1 \geq 0$ and $a_2 \geq 0$. Also from equations (7) and (8) we have

2) $$D_P(x, y) = D'_P(y, x) ,$$

a symmetric condition, and equations (9) and (10) yield

3) $$D_P(x, y) \leq D_P(x, r) + D'_P(y, r) ,$$

a triangle inequality. A function satisfying such conditions will be termed a "polarized metric."

V. EXCLUSION REGION

Now consider the process of keeping point x fixed and letting point y vary. Of special interest are the points y close enought to x such that the C and D regions of x overlap with the C´ and D´ regions of y. Define region E_1 as the set of points y such that the C and C´ regions overlap, region E_2 as the set of points y such that the D and C´ regions overlap, and region E_3 as the set of points y such that the C and D´ regions overlap. Define the exclusion-region or E-region as the union of the three regions E_1, E_2, and E_3. These definitions are described more concisely by:

$$y \in E_1 \quad \text{iff} \quad C \cap C' \neq \emptyset ,$$
$$y \in E_2 \quad \text{iff} \quad D \cap C' \neq \emptyset ,$$
$$y \in E_3 \quad \text{iff} \quad C \cap D' \neq \emptyset ,$$

and $$E = E_1 \cup E_2 \cup E_3$$

where $\emptyset$ is the empty set. An example using specific C and D regions is shown in Figure 2.

The importance of this construction is that if no pair of code words of a binary block code has a polarized distance that is contained in the exclusion-region, then the code is capable of correcting all error patterns in the correction-region and detecting all error patterns in the detection-region. This result is stated formally in the following theorem.

Theorem 1: A code can simultaneously correct all errors e such that $W_P(e) \in$ C-region and detect all errors such that

$W_P(e) \varepsilon$ D-region if and only if, for all distinct code words x and y,

$$D_P(x, y) \not\varepsilon \text{ E-region} . \qquad (11)$$

Proof: Both the sufficiency and necessity of condition (11) must be shown.

Sufficiency: It will be shown that if a code is not capable of the specified error correction and detection, then it is impossible for condition (11) to hold. In any such code there exists two code words x and y and an element of the noncode space r such that:

1) $D_P(x, r) \varepsilon C$ and $D_P'(y, r) \varepsilon C'$

or 2) $D_P(x, r) \varepsilon D$ and $D_P'(y, r) \varepsilon C'$

or 3) $D_P(x, r) \varepsilon C$ and $D_P'(y, r) \varepsilon D'$.

In any one of the three cases, r is a correctable error of one code word and a correctable or detectable error of the other, and hence a decoding mistake could be made. Consider equations (9) and (10),

$$d_{10}(x, y) = d_{10}(x, r) + d_{01}'(y, r) - \delta$$

and $$d_{01}(x, y) = d_{01}(x, r) + d_{10}'(y, r) - \delta .$$

The first inequality constrains the horizontal distance between x and y in the polarized distance space and the second inequality constrains the vertical distance. If $\delta = 0$, then in case (1) above, x and y are close enough together that C and C´ overlap and, by definition, $D_P(x, y) \varepsilon E_1$. Similarly, in case (2), $D_P(x, y) \varepsilon E_2$ and in case (3), $D_P(x, y) \varepsilon E_3$. If $\delta \neq 0$, then y moves closer to x along the diagonal shadow line of the point where it would be located for $\delta = 0$. Since the union of the C and D regions is a connected shaded region containing point (0, 0), regions that overlapped when $\delta = 0$ overlap all the more when $\delta \neq 0$ and again $D_P(x, y) \varepsilon E$. This clearly contradicts condition (11). Thus, if a code satisfies condition (11), it is capable of the specified error correction and detection.

Necessity: It will be shown that if $D_P(x, y) \varepsilon E$ for at least one pair of code words, then there exists an element of the noncode space r such that one of the cases (1), (2), or (3) above is true. In attempting to find such an r, first consider only those vectors that contain the same bit values as x and y in those bit positions where x and y agree. In other words,

$$d_{101}(x, r, y) = 0$$

and $$d_{010}(x, r, y) = 0 .$$

From equation (4), this implies that $\delta = 0$. Thus the equations (9) and (10) become

$$d_{10}(x, y) = d_{10}(x, r) + d'_{01}(y, r)$$

and

$$d_{01}(x, y) = d_{01}(x, r) + d'_{10}(y, r)$$

for the elements r under consideration. By the construction of region E, $D_P(x, y) \varepsilon E$ means that C and C´ overlap, or D and C´ overlap, or C and D´ overlap. Since the two equations above are equalities ans do not contain δ, and since they describe horizontal and vertical distances respectively between x, r, and y, it is a simple matter to select an r from those for which $\delta = 0$ that satisfies one of the cases (1), (2), or (3). Thus, any code capable of the specified error correction and detection satisfies condition (11).

VI. SPECIAL CASES

It is fairly obvious from the definition of polarized distance that it is a generalization of the Hamming distance concept [5]. In fact the Hamming weight, $W_H(e)$, of an error e is

$$W_H(e) = w^+(e) + w^-(e) ,$$

the number of nonzero components. Since the Hamming distance, $D_H(x, y)$, between two code words x and y is

$$D_H(x, y) = W_H(x - y) = d_{10}(x, y) + d_{01}(x, y) ,$$

points y that are a constant Hamming distance from x lie on a diagonal line in the polarized distance space and, as shown in Figure 3, errors such that $W_H(e) \leq t$ lie in a triangular region of the polarized distance space. Figure 3 is an illustration of the ideas expressed in the following corollary using the polarized distance technique. Note that the E-region is also triangular.

Corollary 1: A code can simultaneously correct all errors e such that $W_H(e) \leq t$ and detect all errors such that $t \leq W_H(e) \leq d$ if and only if, for all distinct code words x and y,

$$D_H(x, y) > t + d .$$

Freiman [3] considered error detection/correction in a specific asymmetric error model which again is a special case of Theorem 1. If a (t_1, t_2)-region is defined to be the set of all points (a_1, a_2) such that $a_1 \leq t_1$ and $a_2 \leq t_2$, then in this terminology, Freiman considered only the cases in which the C and D regions were of this type. A (t_1, t_2)-region forms a rectangular region in the polarized distance space as shown in Figure 4. The general shape of the E-region is illustrated in Figure 4 for a particular example in which the C-region is a (t_1, t_2)-region and the D-region is a (d_1, d_2)-region. This idea is expressed formally in the following corollary:

Corollary 2: A code can simultaneously correct all errors e such that $W_P(e) \leq (t_1, t_2)$ and detect all errors such that $(t_1, t_2) \leq W_P(e) \leq (d_1, d_2)$ if and only if, for all distinct code words x and y,

$$\max[d_{10}(x, y), d_{01}(x, y)] > \max(t_1 + d_2, t_2 + d_1)$$

or $$\min[d_{10}(x, y), d_{01}(x, y)] > \min(t_1 + d_2, t_2 + d_1) .$$

Freiman proved the above theorem in [3] but erroneously implied that when $t_1 = t_2$ and $d_1 = d_2$, the Hamming symmetric error case resulted. This implication is incorrect since the E region in the Hamming case is always triangular and when $t_1 = t_2$ and $d_1 = d_2$ in the Freiman case, the region is square. This can be seen in Figure 4 by noting that $\min(t_1 + d_2, t_2 + d_1) = \max(t_1 + d_2, t_2 + d_1)$ in this case.

The polarized distance conept is a useful tool for analyzing the capabilities of binary codes in cases where it is important to distinguish between 0-errors and 1-errors. However, the polarized distance concept does have certain important limitations since it only considers the number of 0-errors and the number of 1-errors and ignores any positional dependency among bit errors. Thus, the polarized distance concept is of little use for analyzing the capabilities of codes in cases where errors are of a burst, serial, or arithmetic nature.

VIII. ASYMMETRIC AND UNIDIRECTIONAL ERRORS

This section will be concerned with the detection and correction of errors with two particular probability distributions. The first type, called asymmetric errors, will refer to the situation when the most likely error patterns occur in the weight classes near one axis of the polarized distance space. An example is shown in Figure 5(a) in which 1-errors are much more likely than 0-errors. The second type, called unidirectional errors, will refer to the situation when the most likely error patterns occur in the weight classes near either axis of the polarized distance space. An example is shown in Figure 5(b) in which 1-errors and 0-errors are both likely, but it is unlikely that they occur at the same time.

Referring again to Figure 5, six specific kinds of error detection and correction are shown, each of which is given a special name. These names and the conditions that codes must satisfy to qualify for that particular classification are given below:

(a) CAED (Completely-Asymmetric-Error-Detection) This is a special case of Corollary 2 in which $t_1 = t_2 = d_2 = 0$ and $d_1 = n$. Therefore, the polarized distance condition between code words is:

$$\min[d_{10}(x, y), d_{01}(x, y)] > 0 .$$

(b) CUED (Completely-Unidirectional-Error-Detection)
This is not a special case of Corollary 2; however, by simple construction of the E-region it is seen that the polarized distance condition is exactly the same as the previous case. Thus any code capable of CAED is also capable of CUED.

(c) SAEC (Single-Asymmetric-Error-Correction)
This again is a special case of Corollary 2 in which $t_1 = d_1 = 1$ and $t_2 = d_2 = 0$. Therefore, the polarized distance condition between code words is:

$$\max[d_{10}(x, y), d_{01}(x, y)] > 1 .$$

Note that a code containing two code words at polarized distance (1, 1) is not capable of even the simplest type of error correction and yet may satisfy the CUED conditions.

(d) SEC (Single-Error-Correction)
This is the same as single-unidirectional-error-correction and is a special case of Corollary 1 in which $t = d = 1$. Therefore, the polarized distance condition between code words is:

$$d_{10}(x, y) + d_{01}(x, y) > 2 .$$

(e) SAEC/CAED
This again is a special case of Corollary 2 in which $t_2 = d_2 = 0$, $t_1 = 1$, and $d_1 = n$. Therefore, the polarized distance condition between code words is:

$$\min[d_{10}(x, y), d_{01}(x, y)] > 1 .$$

(f) SEC/CUED
This is not a special case of Corollary 2; however, by simple construction of the E-region it is seen that the polarized distance condition is exactly the same as the previous case. Thus any code capable of SAEC/CAED is also capable of SEC/CUED.

VIII. CAPABILITIES OF SOME SPECIFIC CODES

After determining the conditions that a code must satisfy for a particular type of error detection/correction, there is still the essential task of finding some codes that satisfy them. The weight-distance concepts do not provide a method for constructing such codes; they only provide a tool for analyzing their capabilities after the codes are defined. Of particular use are those codes that not only have desirable capabilities, but also have fast simple decoding algorithms and high rate. Of codes that satisfy the previous SEC conditions, the familiar Hamming code [5] has the highest rate of any code for values of $n = 2^r - 1$, where r is the number of check bits. The Hamming code also has the highest rate of any linear SEC code for any value of n. However, other codes may have simpler decoding algorithms for particular implementations. Kim and

Freiman [9] discussed a nonlinear code that met the SAEC conditions and had higher rate than a Hamming code for any value of n.

Codes that satisfy the CUED conditions (or equivalently the CAED conditions) are particularly useful in the design of self-checking circuits since unidirectional errors are the most likely type of errors in many digital circuits. It is known that the m-out-of-n or fixed-weight codes, in which each code word contains the same number m of 1s, satisfy the CUED conditions. This is readily seen since the polarized distances between code words all lie on the $d_{10} = d_{01}$ line such as (1, 1), (2, 2), etc. Since all fixed-weight codes have two code words at polarized distance (1, 1), they are not capable of error correction.

An m-out-of-n code has $C^n_m = n!/m!(n - m)!$ code words. For fixed n, this number is a maximum for m = [n/2], the integer part of n/2. It is also known that there is no code satisfying the CUED conditions that has a higher rate for fixed n than the [n/2]-out-of-n code. One drawback of the fixed-weight codes is that they are nonseparable. In many practical situations, it is required that a separable code be used in order that k bits of information may be represented arbitrarily. The size of such codes is thus 2^k. Berger [2] described a separable code that satisfies the CUED conditions. In the Berger code the check bits form a binary number equal to the number of 0s in the k information bits. (An equivalent code in terms of the number of check bits is one in which the check bits are the complement of the number of 1s in the k information bits.) It can be seen that the Berger code satisfies the CUED conditions, by noting that an error changing only 0s to 1s in a code word can only increase the binary number formed by the check bits and decrease the number of 0s in the information bits. Thus such an error cannot produce another code word. Similarly, errors changing only 1s to 0s cannot produce another code word. Freiman [3] showed that no other separable CUED code has higher rate than the Berger code.

Note that a Berger code is no more redundant than a Hamming SEC code and, for many values of n, is less redundant. The [n/2]-out-of-n code is less redundant than the Berger code and, as shown by Freiman [3], has effectively one-half as many check bits as a Berger code for large values of n. Thus, an [n/2]-out-of-n code is also less redundant than a Hamming SEC code for all values of n.

Kautz and Elspas [8] considered a new code in which code words consist of a constant-weight data portion and additional check bits. They presented a method for generating the check bits to produce a SEC code. This code was analyzed using the polarized distance technique and was found to satisfy the SEC conditions. However, such codes have two code words at polarized distance (2, 1) and, therefore, are not SEC/CUED codes.

IX. CONCLUSIONS

A polarized distance technique was introduced and shown to be useful for determining the error-detecting/correcting capabilities of binary block codes for nonrandom errors. The special cases of asymmetric and unidirectional error detection and correction were considered in detail. Using the polarized distance technique, conditions for the detection and correction of such errors were determined and several specific codes were shown to satisfy some of these polarized distance conditions.

ACKNOWLEDGMENT

This research was supported in part by an NDEA Fellowship under Title IV, in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under contract DAAB-07-67-C-0199, and in part by the National Science Foundation under grant GK-15459. This work was performed at the Coordinated Science Laboratory, University of Illinois, Urbana, and appears in Chapter 2 of the author's Ph.D. thesis [1].

REFERENCES

1. Anderson, D. A., "Design of Self-Checking Digital Networks Using Coding Techniques," Coordinated Science Lab Report R-527, Univ. of Illinois, Ph.D. Thesis, October 1971.

2. Berger, J. M., "A Note on Error Detection Codes for Asymmetric Channels," Information and Control, Vol. 4, pp. 68-73, March 1961.

3. Freiman, C. V., "Protective Block Codes for Asymmetric Binary Channels," Columbia Univ., Ph.D. Thesis, May 1961.

4. Garner, H. L., "Generalized Parity Checking," IRE Trans. on Electronic Computers, Vol. EC-7, pp. 207-213, September 1958.

5. Hamming, R. W., "Error Detecting and Error Correcting Codes," Bell System Tech. Journal, Vol. 29, pp. 147-160, April 1950.

6. Hsiao, M. Y. and Tou, J. T., "Application of Error-Correcting Codes in Computer Reliability Studies," IEEE Trans. on Reliability, Vol. R-18, pp. 108-118, August 1969.

7. Kautz, W. H., "Codes and Coding Circuitry for Automatic Error Correction within Digital Systems," Redundancy Techniques for Computing Systems, Wilcox and Mann, eds., Spartan Books, Washington, D.C., pp. 152-195, 1962.

8. Kautz, W. H. and Elspas, B., "Single-Error-Correcting Codes for Constant-Weight Data Words," IEEE Trans. on Information Theory, Vol. IT-11, pp. 132-141, January 1965.

9. Kim, W. H. and Freiman, C. V., "Single Error-Correcting Codes for Asymmetric Binary Channels," IRE Trans. on Information Theory, Vol. IT-5, pp. 62-66, June 1959.

10. Kim, W. H. and Freiman, C. C., "Multierror Correcting Codes for a Binary Asymmetric Channel," IRE Trans. on Information Theory, Vol. IT-5, Special Supplement, pp. 71-78, May 1959.

11. Marcus, M. P., "Minimum Polarized Distance Codes," IBM Journal of Research and Development, Vol. 5, pp. 241-248, July 1961.

12. Massey, J. L., "Survey of Residue Coding for Arithmetic Errors," ICC Bulletin, Vol. 3, Rome, Italy, pp. 195-209, October 1964.

13. Peterson, W. W., Error-Correcting Codes, The M.I.T. Press, Cambridge, 1961.

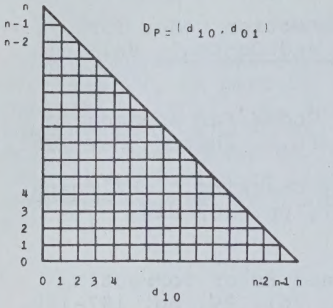

n
n-1
n-2
$D_P = (d_{10}, d_{01})$
d_{01}
4
3
2
1
0
0 1 2 3 4
n-2 n-1 n
d_{10}

FIGURE 1 : POLARIZED DISTANCE SPACE

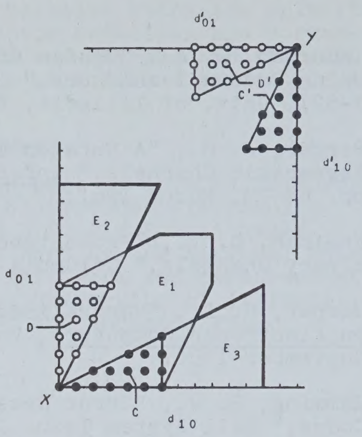

d'_{01}
Y
D'
C'
d'_{10}
E_2
d_{01}
E_1
D
E_3
X
C
d_{10}

FIGURE 2 : EXAMPLE C,D,C',D', AND E REGIONS

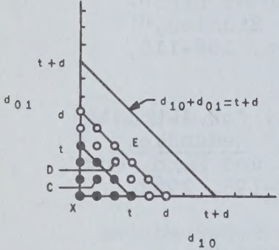

t + d
d_{01}
$d_{10}+d_{01}=t+d$
d
E
t
D
C
X
t
d
t + d
d_{10}

FIGURE 3 : SYMMETRIC ERROR DETECTION AND CORRECTION

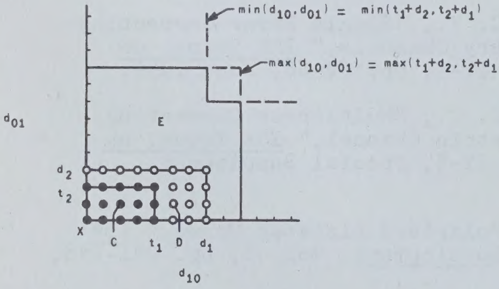

$\min(d_{10}, d_{01}) = \min(t_1+d_2, t_2+d_1)$
$\max(d_{10}, d_{01}) = \max(t_1+d_2, t_2+d_1)$
d_{01}
E
d_2
t_2
X
C
t_1
D
d_1
d_{10}

FIGURE 4 : ASYMMETRIC ERROR DETECTION AND CORRECTION

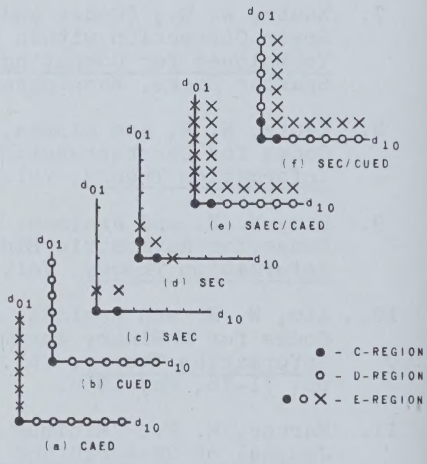

d_{01}
d_{10}
(f) SEC/CUED
d_{01}
d_{10}
(e) SAEC/CAED
d_{01}
d_{10}
(d) SEC
d_{01}
d_{10}
(c) SAEC
d_{01}
d_{10}
(b) CUED
d_{01}
d_{10}
(a) CAED
● - C-REGION
○ - D-REGION
● ○ ✕ - E-REGION

FIGURE 5 : ASYMMETRIC AND UNIDIRECTIONAL ERROR DETECTION AND CORRECTION

SYSTEMATIC NON-SEPARATE ARITHMETIC GROUP CODES

OSCAR N. GARCIA and ROBERT T. GREENE
College of Engineering, University of South Florida
Tampa, Florida 33620

ABSTRACT

Arithmetic group codes are defined in the context of finite groups where the addition operation is defined by a type of directed carry-flow graph called spined. The vertices of such graphs have associated weights which are a possible solution of a system of congruences. Using these concepts the arithmetic codes are classified, and the systematic non-separate codes are considered in some detail. The addition modulo m in a radix r representation is considered with view to its simpler logic realization. The results obtained are applied to a class of error detecting and correcting systematic group codes.

INTRODUCTION AND BACKGROUND

Arithmetic codes in the infinite ring of integers have been studied by Henderson[1], Peterson[2], Chien[3], Garner[4], Mikhelev[5], and by Massey[6] among others. A well known class of arithmetic codes called AN codes, have been treated as subsets of the radix r (most often $r = 2$) expansion of the integers subject to the constraint that the expansion of the product of a constant A times an information integer N is a valid result of the integer addition of two other code words AN_1 and AN_2 only if $N_1 + N_2$ is less than a certain positive integer M, called the range of the code. This restriction is simply removed in the following:

Definition: An arithmetic group code C is a subgroup of a group formed by a certain subset of all the n-tuples with elements from $S_r = \{0, 1, 2, \ldots, r-1\}$ with $1 < r < m$ and an operation so defined that the group is isomorphic to Z_m, the integers under addition modulo m.

This definition allows these codes to be considered as finite additive number systems, and the isomorphic mapping to be an interpretation function[7].

This isomorphic mapping (call it $\alpha: C \to Z_m$) may be specified in a variety of ways. We will choose to further specialize the codes as defined above, to those in which it is possible to establish a congruential relation modulo m between the code word and the corresponding element of Z_m. This is done by assigning a weight w_i to each digit position i of the n-tuple $X_j = [x_{j,n-1}, x_{j,n-2}, \ldots, x_{j,0}]$ representing a code word and then requiring that

$$\sum_{i=0}^{n-1} x_{j,i} w_i \equiv \alpha(X_j) \bmod m \qquad (1)$$

for all code words, i.e., $j = 0, 1, \ldots, m-1$. Given a mapping α as described, the question of whether such a congruential relation may be established or not will depend on whether or not the system of m linear congruences (1) is solvable. This interesting question is analogous, but not identical, to the solution of systems of linear equations. Let us just point out here that it is known, for example, that given n linearly independent congruences in n indeterminates there is a unique solution if the determinant of the matrix of coefficients is relatively prime to the modulus. If this is not the case then the system of congruences for $m = \Pi p_i^{q_i}$ and p_i prime, will be solvable if it is solvable for each modulus $p_i^{q_i}$ separately, and if there are P_i solutions for that modulus then the total number of solutions will be ΠP_i. To find P_i the reader is referred to Smith[8].

We will assume now that for the codes that we will deal with in this paper a set of weights w_i for $i = 0, 1, \ldots, n-1$ solves this system of congruences. The codes may be visualized now as finite additive weighted number systems, usually redundant so that their distance properties may be utilized for error correction.

Carry Arithmetic Representation by a Graph

Once we have the n-tuples and their interpretation we could just have a table look-up procedure to determine how any two combine arithmetically. A much simpler procedure is to specify, when possible, how to combine arithmetically the n-tuples by a directed graph showing the flow of carries.

Example 1: Let $r = 2$, $n = 3$ and $m = 5$, with the set of triples {111, 010, 011, 101, 110} and their interpretation into Z_5 being 1, 2, 3, 4, 0 respectively. Their carry structure may be represented as in Figure 1.

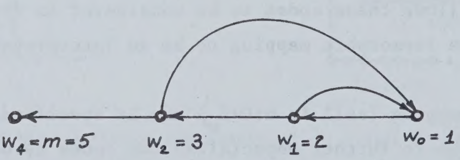


Figure 1. Carry-Flow Graph for Example 1

Whenever we have corresponding bit positions which add to r or more, then a carry is generated into the other digit positions connected by outgoing branches from the corresponding vertex in the carry-flow graph. Notice that these carries may generate new carries.

If in our Example 1 we add (011) + (101) we get a modulo 2 bitwise sum of 110 and a carry of 010 which when added again generates a mod 2 sum 100 and a carry of 101; these two added yield 001 and a carry of 001 which result in the final sum 010. Notice also that the vertex with weight 5 acts like a sink, i.e., carries disappear into it.

Out of all the possible carry-flow graphs we will specialize in those in which each vertex is connected by a single branch to another adjacent vertex, say to the left, and to possibly others to the right. The precise definition of this type of graph, called spined, and more details on these representations have been given by Garcia[9]. The graph in Figure 1, for example, is spined.

The Classification of Arithmetic Codes

Now we can classify all arithmetic codes as the following types: separate codes, which are those in which two or more not connected spined carry-flow graphs are needed to represent the arithmetic of system, and non-separate which are those for which a single spined carry-flow graph suffices. Examples of separate arithmetic codes are the multi-residue codes studied by Rao and Garcia[10]. Non-separate codes are for example the AN codes, which have been known for some time. For a review of the AN codes from a finite arithmetic point of view see Massey and Garcia[11]

If we consider now the encoding of k-tuples, representing integers from 0 to M-1, into n-tuples we can classify the codes as systematic, for those in which the digits of the k-tuples are directly identifiable in fixed positions of their encoded form, and non-systematic, for those in which they can't be recognized. The multi-residue codes[10], for example, are always systematic. The AN codes are, in general, non-systematic but we will see how it is possible to obtain systematic non-separate codes.

If using the n-tuple $X = (x_{n-1}, x_{n-2}, \ldots, x_0)$ where $x_i = 0$ for $i = 1, 2, \ldots, n-1$ and $x_0 = 1$ we generate a unary algebra[12] by adding X to itself and to the successive sums according to the rules of a spined graph, the transition graph of the system would look, in general, as in Figure 2.

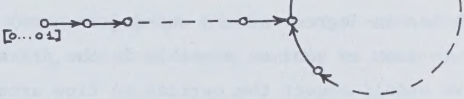


Figure 2. Transition Graph of Unary Algebra

The n-tuples so generated form a semi-group under addition as defined by the spined graph; the circular portion of the transition graph forms a group, and a subgroup of this group is an arithmetic code.

SYSTEMATIC NON-SEPARATE CODES

Systematic arithmetic codes have been studied in particular by Henderson[1], Garner[4], Mikhelev[5], Rao[13] and Trehan[14]. Henderson[1] was the first one to point to the systematic nature of a class of arithmetic codes. Garner[4] and Mikhelev[5], independently, pointed out that given an AN code, not necessarily systematic, it can be recoded by a permutation of the coding function α, when the generator $A = a \cdot m'$, $a > 1$ and m' is relatively prime to M by solving the Diophantine equation

$$am' + kM = 1 \tag{2}$$

where m' and M are relatively prime.

We see here that in terms of our directed graph representation the general form for the arithmetic of these codes is given in Figure 3, with $m = m' \cdot M$[15,16].

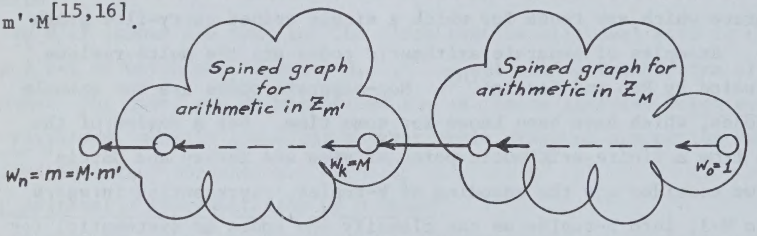

Spined graph for arithmetic in $Z_{m'}$
Spined graph for arithmetic in Z_M
$W_n = m = M \cdot m'$
$W_k = M$
$W_0 = 1$

Figure 3. Di-graph for General Carry Flows in Systematic Non-separate Codes

Rao[13] has given a definition for systematic distance, which has been used by Trehan[14] to determine the error correcting properties of systematic non-separate codes.

The Arithmetic of a Class of Systematic Non-separate Codes

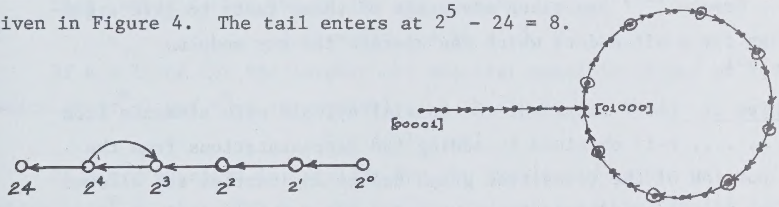
One of the problems that we face in the implementation of the arithmetic as defined by a spined graph, among the elements of the circular portion (group) of the transition graph in Figure 2, is that some of the generated carries generate other carries which in turn could generate other carries and so on. We are guaranteed that this process ends if the modulus of the system (i.e., the weight associated with the unique vertex of the spined graph which has in-degree one and out-degree zero) is positive. But we would like to know as soon as possible in the process of addition how many times we should expect the carries to flow around before settling to the final result. This is possible for systematic non-separate codes in which $M = r^k$ and $m = m' \cdot r^k$. The following development although more general, applies to this case, and has been considered by Greene[17] with regards to the implementation of modular addition with $r = 2$.

Consider in what follows a spined graph for which $w_i = r^i$ for $i = 0, 1, \ldots, n-1$ and where the weight $w_n = m$ associated with the modulus vertex satisfies $r^n \geq m > r^{n-1}$. Then for any n-tuple $A = [a_{n-1}, a_{n-2}, \ldots, a_0]$ define the integer valued function $\delta(A) = \sum_{i=0}^{n-1} a_i w_i$ where w_i is the weight associated with the $i^{\underline{th}}$ vertex.

Lemma: The point where the tail enters the circular portion of the transition graph (Figure 2) corresponds to the binary n-tuple X for which $\delta(X) = r^n - m$ for a system with a modulus $m < r^n$. For $m = r^n$ the transition graph is a circle without a tail.

Proof: Under the assumptions given all r^n-1 non-zero n-tuple representations will occur. When $N_1 = [0, 0, \ldots, 0, 1]$ is added to $N_2 = [r-1, r-1, \ldots, r-1]$ we will obtain a carry into the modulus vertex and carries to other vertices with $w_i < r^{n-1}$ such that their sum is $r^n - m$. If $m = r^n$ we have an all-zero representation, and all r^n n-tuples lie in a circle without a tail. Otherwise, the representation for r^n-m must have occurred before and the circle closes exactly on it.

Example 2. Let $r = 2$, $m = 24$, $M = 8$. We could consider the single error detecting systematic group code {01001, 10010, 11011, 01100, 10101, 11110, 01111, 11000} which is obtained by using the encoding function 9N for $0 < N \leq 8$ where, of course, $N = 8 \equiv 0$. This is the solution of (2) with $a = 3$, $m' = 3$, $k = -1$ and $M = 8$. The spined and transition graphs are given in Figure 4. The tail enters at $2^5 - 24 = 8$.



(a) Spined graph (b) Transition graph

Figure 4. Graphs Corresponding to Example 2

Theorem 1: Addition of two code words (i.e., representations from the circular portions of the transition graph of Figure 2) according to the rules of the spined graph yields another code word (closure). Furthermore, if N_1 and N_2 are code words, and N_3 is their sum modulo m then

$$\delta(N_1) + \delta(N_2) = \delta(N_3) + km \tag{3}$$

where the positive integer $k < r+1$ and m is the modulus of the number system.

Proof: The closure property is a direct result of the fact that any representation N_3 in the circle is the result of adding [0 ... 01] exactly $\delta(N_2)$ times to N_1; since we start with a representation in the circular portion of the transition graph, the addition of [0 ... 01] to it any number of times yields a representation in the circle also.

Let k_{max} be the largest value of k in (3); the largest values of $\delta(N_1) = \delta(N_2) = r^n-1$ and the smallest of $\delta(N_3) = r^n-m$ (by the lemma above) will yield the largest product of km. Solving under these assumptions in (3)

$$k \leq k_{max} = \frac{r^n-2}{m} + 1$$

but the smallest m can be is $r^{n-1}+1$ under our assumptions. Then

$$k \leq \frac{(r+1)r^{n-1}-1}{r^{n-1}+1} = (r+1) - \frac{r+2}{r^{n-1}+1} < r+1$$

which proves our assertion.

The implication of this result is the assurance that for any modulus under the given conditions the carries will not flow around more than twice. But we can do better than that, as we shall see below, by predicting even before the carries that flow out of the $n^{\underline{th}}$ bit position are added again, what value of k we will have. This means that we do not actually have to wait for the flow of these carries again, but we can tell whether r^n-m or $2(r^n-m)$ will have to be added to the interim sum and once this is done no new carries will flow out of the most significant digit (m.s.d.). Greene[17] has taken advantage of these facts to give a general design for n-bit adders which can operate for any modulus $2^n \geq m > 2^{n-1}$.

Theorem 2: Let P represent the initial n-tuple with elements from $S_r = \{0, 1, ..., r-1\}$ obtained by adding two representations from the circular portion of the transition graph before any carries are allowed to flow out of the m.s.d. and let k be as in (3). Then only the following cases are possible: (for r=2)

1) If $k = 0$ then $2(r^n-m) \leq \delta(P) \leq r^n-1$

2) If $k \neq 0$ then either: a) $0 \leq \delta(P) \leq m-1$ and $k = 1$, or

b) $m \leq \delta(P) \leq r^n-2$ and $k = 2$.

Proof: Before we proceed let's point to the transition graph of Figure 5 where each one of the areas where the integer value of the interim sum P is located. It is clear also that immediately after P is formed in a standard carry ripple adder, it is known whether k = 0 or not. Also, we have included for completeness the possibility of an all-zero representation in the graph.

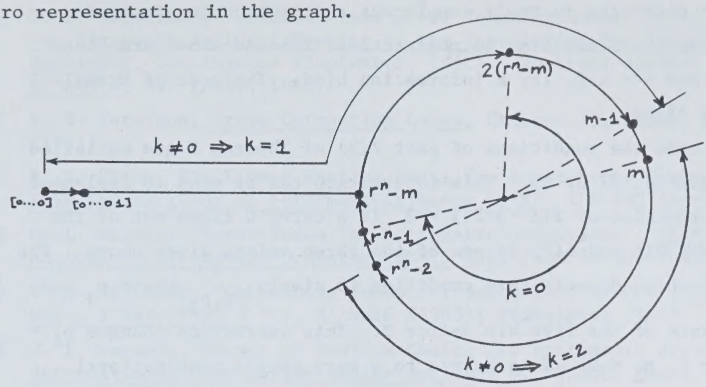


Figure 5. The Three Areas in the Transition Graph Shown by Values of $\delta(P)$

If $k = 0$ in (3), since $\delta(P) = \delta(N_3)$ in this case, then

$$\delta(P) = \delta(N_1) + \delta(N_2)$$

But the largest $\delta(N_3)$ can be is r^n-1 and the smallest $\delta(N_1) + \delta(N_2)$ can be is $2(r^n-m)$. Then

$$2(r^n-m) \leq \delta(P) \leq r^n-1$$

If $k \neq 0$ we can say that

$$\delta(N_1) + \delta(N_2) = \delta(P) + r^n \tag{4}$$

where r^n is the total numeric value of the branches flowing out of the vertex with weight r^{n-1} in the spined graph.

Then, substituting (4) in (3) and solving

$$\delta(P) = \delta(N_3) + km - r^n \tag{5}$$

If $k = 1$ and for the largest and smallest possible values of $\delta(N_3)$ which are r^n-1 and r^n-m respectively, we obtain in (5) that

$$0 \leq \delta(P) \leq m-1$$

If $k = 2$ equation (5) yields, for the smallest possible value of $\delta(N_3)$ which is r^n-m, that $\delta(P) \geq m$ and for the largest possible $\delta(N_1) + \delta(N_2)$ which is $2(r^n-1)$ equation (4) gives that $\delta(P) \leq r^n-2$.

Since the bounds given for $k \neq 0$ are not overlapping, and no other cases are possible, the theorem is proved.

Two interesting particular cases of this theorem give known results in computer arithmetic.

Corollary: For radix complement arithmetic ($m = r^n$) and diminished radix complement arithmetic ($m = r^n-1$) we will have $0 \leq k < 2$.

APPLICATIONS AND CLOSING REMARKS

For error detection in two's complement arithmetic the simplest systematic codes which satisfy the restrictions imposed above are those with $m = 3 \cdot 2^k$ and $n = k+2$, for k information bits. The code of Example 2 is one of such class.

For that code the conditions of part 2(b) of Theorem 2 are satisfied when $\delta(P)$ equals 24, 27 or 30. This information can be used to implement an additive correction of $2(2^5 - 3 \cdot 2^3) = 2^4$ if a carry C flows out of the most significant bit and $\delta(P)$ is one of the three values given above. The logic function which detects this condition is simply Cp_4p_3 where p_4 and p_3 are components of the five bit vector P. This correction changes $p_4 = p_3 = 1$ to $p_4 = 1$, $p_3 = 0$. This points to a very simple combinational implementation of an otherwise complex sequential ripple-carry flow. It is, of course, also possible to do the same for condition 2(a) of Theorem 2. The application of the results of Theorem 2 have been illustrated with a very simple example with $m' = 3$. They can be useful when the values of m' are such that finding a combinational function to implement the correction is difficult.

For binary arithmetic error correction using again $M = 2^n$ in (2) we can choose $m' > 2n+1$ such that if x is the smallest positive integer for which $2^x \equiv \pm 1$ modulo m' then $n \leq x$. (This integer x has been called the length of the code[10] modulo m'.). This guarantees distinctiveness of the residues of the errors in any of the n bits of the code. Similarly m' could be chosen an odd prime $p > n$ which has 2 (or -2) as a primitive root as used by Peterson[2], or the product of two primes as used by Rao and Garcia[10] with proper adjustment for the length of the code. For an arbitrary radix r, a generalization of the result of Peterson[10] mentioned above due to Rao and Trehan[18], could also be useful in choosing a value for m'.

Finally, the reader should be reminded that for these codes in arithmetic modulo m the generator is $A = a \cdot m'$ where the coefficient a is obtained from the solution of (2) under the given assumptions.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by the National Science Foundation Research Initiation Grant GK-25278 at the University of South Florida.

REFERENCES

[1] D. S. Henderson, "Residue Class Error Checking Codes", Proceedings of the Sixteenth National Meeting of the Association for Computing Machinery, Los Angeles (September, 1961). Abstract appears in Comm. of ACM 4, 307 (July, 1961).

[2] W. W. Peterson, Error Correcting Codes, Chapter 13, Wiley, New York (1961).

[3] R. T. Chien, "On Linear Residue Codes for Burst Error Corrections", IEEE Transactions on Information Theory IT-IV, 126-133 (April, 1964).

[4] H. L. Garner, "Error Codes for Arithmetic Operations", IEEE Transactions on Electronic Computers EC-15 (5), 763-770 (October, 1966).

[5] V. M. Mikhelev, "Correcting Codes for Arithmetic Errors", Zh. Vych. Mat., i Mat. Fiz. 5 (2), 311-316 (1965); Translated (1965).

[6] J. L. Massey, "Survey of Residue Coding for Arithmetic Errors", International Computation Center Bulletin 3, UNESCO, Rome, Italy, (October, 1964).

[7] H. L. Garner, "The Classification of Finite Number Systems", IFIP, Congress Proceedings, Edinburg, Scotland (August 5-10, 1968).

[8] H. J. S. Smith, Collected Mathematical Papers, Paper V, Vol. I, pp. 43-5, Chelsea Publishing Co., New York, (reprinted) (1965).

[9] O. N. Garcia, "On Weighted Representations for Modular Arithmetic", Fifth Annual Princeton Conference on Information Sciences and Systems, Princeton, New Jersey, (March 25-26, 1971).

[10] T. R. N. Rao and O. N. Garcia "Cyclic and Multi-residue Codes for Arithmetic Operations", IEEE Transactions on Information Theory IT-17, 85-91 (January, 1971).

[11] J. L. Massey and O. N. Garcia, "Error Correcting Codes for Computer Arithmetic", in Advances in Information Systems Science, Vol. IV, Plenum Press (to be published).

[12] G. Birkhoff and T. C. Bartee, Modern Applied Algebra, Chapter 2, Section 2-7, McGraw-Hill, New York (1970).

[13] T. R. N. Rao, "On Systematic Non-separate Arithmetic Codes", Proceedings of the Fourth Annual Princeton Conference on Information Sciences and Systems, Princeton, New Jersey (March, 1970).

[14] A. K. Trehan, "Residue Codes and Their Application to Arithmetic Processors", Ph.D. Dissertation, Department of Engineering, University of Maryland, 1971.

[15] O. N. Garcia, "Graphs, Number Systems and Arithmetic Codes", NASA Document X-711-69-384, Goddard Space Flight Center, Greenbelt, Md. (October, 1969).

[16] O. N. Garcia, "Notes on Arithmetic Codes", Report # TR-EGE-1-71, Department of Electrical and Electronic Systems, College of Engineering, University of South Florida, Tampa, Florida, Section 5 (January, 1971).

[17] R. T. Greene, "The Design of Modular Adders", M.S. Thesis, Department of Electrical and Electronic Systems, College of Engineering, University of South Florida, Tampa, Florida (submitted, 1971).

[18] T. R. N. Rao and A. K. Trehan, "Single-Error-Correcting Nonbinary Arithmetic Codes", IEEE Transactions On Information Theory IT-16 (5), pp. 604-608 (September, 1970).

POWER SYSTEM CONTROL PRACTICE

by

Nathan Cohn
Leeds & Northrup Company
North Wales, Pennsylvania

ABSTRACT

Hierarchical regulating requirements of an interconnected system and its constituent areas are reviewed. Self-regulating elements and their relationships to supplementary controls are identified. Criteria for cooperative inter-area regulation, techniques for achieving energy balances and synchronous time correction, and coordination of area economic and security dispatch with inter-area obligations are discussed.

INTRODUCTION

On multi-area interconnected power systems[1], like the coast-to-coast network that encompasses most of the United States and parts of Canada, there are operating objectives for the system as a whole, and additional objectives for each of the constituent control areas. For the overall system, the objectives, in priority rank, are: (1) $G = L$; (2) $F = F_o$; and (3) $\varepsilon = 0$. (See symbology definitions in section that follows.) For each area they are: (1) Maintain $T_n = T_{on}$ when $F = F_o$; (2) Cause T_n to depart from T_{on} in programmed extent to assist areas in need when $F \neq F_o$; (3) Share in achieving $F = F_o$; (4) Share in achieving $\varepsilon = 0$; (5) Reduce I_n to 0; and (6) Optimize area economy consistent with security constraints.

This paper identifies and defines as primary regulation the self-regulating effects present in all areas which, without additional control, fulfill the major system objective of achieving $G = L$. It then defines the nature of secondary or supplementary controls by means of which remaining system objectives and the tabulated area objectives are achieved. Criteria for cooperative inter-area regulation are reviewed, the concept of automatic adaptive shifts of bulk power transfer schedules is introduced, techniques for achieving energy balance and synchronous time correction are described, and coordination of area economic and security dispatch with inter-area obligations briefly discussed. Throughout the paper the emphasis is on concepts and techniques for establishing control "set points" which, if achieved, would yield desired system and area operating objectives, rather than on details of control system execution. References are provided for those who may wish to pursue more fully the latter aspects of power system control practice.

SYMBOLOGY AND NOTATIONS

The interconnected system considered in this paper will be assumed to consist of N areas. Each area n has J_n generating units, and has interconnecting tie lines with M_n other areas. Unless otherwise stated, all frequency parameters are in Hz, and all power parameters in megawatts.

Parameters Related to System Frequency and Time Deviation

F_r is the rated system frequency; 60 Hz in the U.S.

F_o is the system frequency schedule.

F is the actual system frequency.

ϕ_{on} is any error or offset in setting system frequency schedule at area n.

ϕ_{1n} is any error in measuring system frequency at area n.

ϕ_n is the total "phi" error at area n, and is given by:

$$\phi_n = \phi_{on} + \phi_{1n} \tag{1}$$

t is the time span in hours over which system frequency deviation is integrated for computation of system time deviation.

ε is the system time deviation in seconds, and is given by:

$$\varepsilon = \frac{3600}{F_r} \int_0^t (F - F_r)dt \tag{2}$$

A point to be noted for this paper is that F is essentially a continually varying parameter, with a period of a few seconds normally within a band, superimposed on its slower sustained variations, of about ±0.01 Hz on the large U.S. interconnection. This is not unexpected considering the changing demands from many millions of customers, the elastic nature of many of the tie lines, and the parallel operation of many hundreds of turbine-generator governors with their varying characteristics and response times. Thus the system objective of $F = F_o$, referred to in the introduction, is not intended to be, nor could it be achieved as, an instantaneous condition. Rather, F_o is intended to serve as the "base" about which F will have its swings. Thus F may be regarded as the "steady state" value of system frequency, and in this same "steady state" sense, frequency will be assumed to be identical throughout the synchronized system[2].

Parameters Related to Power Generation and Load

$G_{n,j}$ is the power output of generator j in area n.

G_n is the total power output of all generators in area n.

G is the total power output of all generators on the system.

$G_{on,j}$ is the power output corresponding to scheduled frequency F_o on the prevailing generation-frequency governing characteristic of generating unit j in area n.

$G_{Fn,j}$ is the power output corresponding to frequency F on the prevailing generation-frequency governing characteristic of generating unit j in area n.

$\beta_{1n,j}$ is the cotangent in MW/0.1 Hz over the frequency span under consideration, of the generation-frequency characteristic, i.e. the governing characteristic, of generating unit j in area n. It is considered to have a minus sign.

β_{1n} is the cotangent of the combined generation-frequency characteristic of all generating units in area n.

β_{2n} is the cotangent in MW/0.1 Hz over the frequency span under consideration of the load-frequency characteristic of all connected loads in area n.

β_n is the cotangent of the overall frequency-response characteristic for area n, and is given by:

$$\beta_n = \beta_{1n} - \beta_{2n} \tag{3}$$

β_2 is the cotangent of the load-frequency characteristic for all connected loads of the system.

L_{on} is the total connected load in area n, rated at F_o including losses.

L_o is the total connected load of the system, rated at F_o inc losses.

L_n is the effective connected load in area n at frequency F inc losses.

L is the effective connected load of the system at frequency F inc losses.

Parameters Related to Area Net Interchange and Inadvertent Interchange Accumulations

$T_{on,m}$ is the scheduled net interchange between area n and neighboring area m. Power out is considered plus.

T_{on} is the scheduled net interchange between area n and all of its M neighbors, and is given by:

$$T_{on} = \sum_{m=1}^{M_n} T_{on,m} \tag{4}$$

$T_{n,m}$ is the net interchange power flow between area n and neighboring area m.

T_n is the net interchange power flow between area n and all of its M neighbors, and is given by:

$$T_n = \sum_{m=1}^{M_n} T_{n,m} \tag{5}$$

τ_{on} is any error or offset in setting area net interchange schedule at area n.

τ_{1n} is any error in measuring area net interchange at area n.

τ_n is the total "tau" error in area n, and is given by:

$$\tau_n = \tau_{1n} + \tau_{on} \tag{6}$$

t' is the time span in hours over which area net interchange deviation is integrated for computation of area inadvertent interchange.

I_n is the inadvertent interchange accumulation in MWh in area n, and is given by:

$$I_n = \int_0^{t'} (T_n - T_{on})dt \tag{7}$$

Parameters Related to Stored Energy

S_{on} is the stored energy in area n corresponding to system frequency F_o.

S_n is the stored energy in area n at frequency F.

$$S_n = S_{on}\left(\frac{F}{F_o}\right)^2 \tag{8}$$

S_o is the system stored energy corresponding to system frequency F_o.

S is the system stored energy at frequency F.

$$S = S_o\left(\frac{F}{F_o}\right)^2 \tag{9}$$

Additional Parameters Related to Primary or Governing Regulation

$P_{n,j}$ is the governor regulating error for unit j in area n. It is the generation change required from the unit for prevailing system frequency to match its output to its prevailing frequency-generation curve. It is given by:

$$P_{n,j} = G_{n,j} - G_{Fn,j} \tag{10}$$

A minus regulating error causes an increase in unit generation.

Additional Parameters Related to Secondary, or Supplementary, Control

B_n is the frequency bias setting for area n, in MW/0.1 Hz, and is a minus quantity.

Y_n is the size ratio of area n, based on relative size of its frequency bias setting. It is given by:

$$Y_n = \frac{B_n}{\sum_{n=1}^{N} B_n} \tag{11}$$

E_n is the area control error of area n. A minus control error means that area generation must be increased.

b is the time deviation bias setting for all areas in Hz per second of time error.

H is the time span in hours over which accumulated area inadvertent interchange is to be corrected. It is common to all areas.

T_{xon} is the net interchange schedule for area n adaptively biased by system frequency deviation.

T_{yon} is the net interchange schedule for area n adaptively biased by frequency deviation and system time deviation.

T_{zon} is the net interchange schedule for area n adaptively biased by frequency deviation, time deviation and area inadvertent interchange accumulation.

G_{dn} is the total generation required in area n. It is given by:

$$G_{dn} = G_n - E_n \tag{12}$$

$G_{dn,j}$ is the generation desired from unit j in area n, as computed by economic and security dispatch.

PRIMARY REGULATION

Primary regulation is made up of three effects, each of which will now be examined.

Change in Stored Energy With Frequency Variations

When there is a mismatch between system generation and load, the system operating equation, illustrated in Figure 1, becomes:

$$G - L = \frac{dS}{dt} \tag{13}$$

Changes in system stored energy are related to changes in system frequency. When F is dropping, corresponding decreases in S represent transient means of augmenting G to satisfy L when G < L. When F is increasing, S increases thereby absorbing excess generation when G > L. For small changes of F, Equation (13) may be written:

$$G - L = \left(\frac{2S_o}{F_o}\right) \frac{dF}{dt} \tag{14}$$

For a 300,000 MVA system, reduction of system stored energy would of itself serve a 50 MW load increase for a period of 6 seconds with a frequency drop of 0.005 Hz[3].

Frequency Coefficient of Connected Load

The system load-frequency characteristic is given by:

$$L = L_o + 10\beta_2(F - F_o) \tag{15}$$

If frequency is changing due to a mismatch between G and L, L will change its effective value both in magnitude and direction until--assuming no other regulating effects--it is in balance with G, and a steady state frequency is reached.

Equation (14) may be rewritten thusly:

$$G - L_o - 10\beta_2(F - F_o) = \left(\frac{2S_o}{F_o}\right) \frac{dF}{dt} \tag{16}$$

For an interconnected system, a typical value for β_2 is 0.4% per 0.1 Hz. On a system with L_o = 250,000 MW, a load increment of 50 MW can be satisfied on a sustained basis without any additional generation change, with a steady state frequency drop of 0.005 Hz.

Generator Governing

Each turbine-generator unit is normally equipped with a speed governor, a feedback controller that automatically alters input to the turbine, and hence output from the generator, when it senses changes in system frequency. The operating characteristic of each such governor, derived from Equation (10), is defined by:

$$P_{n,j} = G_{n,j} - G_{on,j} - 10\beta_{1n,j}(F - F_o) \tag{17}$$

The $\beta_{1n,j}$ factors are by no means constant, varying with generator loading and the magnitude of frequency changes. Individual generator-frequency characteristics are therefore not linear. They may have deadbands and hysteresis, and for small frequency changes $\beta_{1n,j}$ may be zero.

We may now write Equation (16) as follows:

$$\sum_{n=1}^{N} \sum_{j=1}^{J_n} [P_{n,j} + G_{on,j} + 10\beta_{1n,j}(F - F_o)] - L_o - 10\beta_2(F - F_o) = \left(\frac{2S_o}{F_o}\right) \frac{dF}{dt} \tag{18}$$

There is combined in Equation (18) all three primary regulating responses. Each governor operates to reduce its $P_{n,j}$ to zero, increasing generation when $P_{n,j}$ is minus. A match is reestablished between G and L, with a corresponding new value for F, when $P_{n,j}$ is zero for all values of n and j. The transient assistance from stored energy is no longer required and $\frac{dF}{dt}$ becomes zero.

Note that whereas the stored energy and load-frequency responses may be considered as instantaneous, the governor control action in reducing $P_{n,j}$ to zero is not. It may take several seconds or more for this to occur at each participating unit. On the large interconnection there may be as many as 3,000 or 4,000 turbine-generators responding at any given time to their respective speed governors.

Primary regulation in accordance with Equation (18) will thus satisfy the overall system objective of matching generation to load. But frequency will be off normal, system time error will develop, and there will be random distribution of total system generation among the participating control areas. Let's see how supplementary control, operating in each of the areas, overcomes these limitations, and fulfills area objectives.

SUPPLEMENTARY CONTROL

The first three area objectives tabulated in the introduction are achieved by operating each area on net interchange tie-line bias control[4], defined by the relationship:

$$E_n = (T_n - T_{on}) - 10B_n(F - F_o) \tag{19}$$

Each area control adjusts area generation by shifting governor speed changers, altering the magnitude of $G_{on,j}$ on one or more area generators to reduce E_n to zero. Generation is increased when E_n has a minus sign.

The $10B_n(F-F_o)$ term of Equation (19) may be regarded as introducing an adaptive shift of area net interchange schedule with frequency deviation, so that Equation (19) may be rewritten:

$$E_n = T_n - T_{xon} \tag{20}$$

where,

$$T_{xon} = T_{on} + 10B_n(F - F_o) \tag{21}$$

It can readily be shown[5] that if several criteria are fulfilled, namely $\Sigma T_{on} = 0$; $\Sigma T_n = 0$; $E_n = 0$ for all values of n, $\tau_n = 0$ for all values of n, and $\phi_n = 0$ for all values of n, then supplementary controls operating in all areas of the interconnected system will yield $F = F_o$, and $T_n = T_{on}$ for all values of n, thereby fulfilling area objectives (1) and (3) tabulated in the introduction.

We will shortly explore the operating results obtained if the foregoing criteria are not fulfilled. First, let's write the generation-load relationship for each area, illustrated in Figure 2, thusly:

$$G_n - L_n - T_n = \frac{dS_n}{dt} \tag{22}$$

Making appropriate substitutions, Equation (22) becomes:

$$\sum_{j=1}^{J_n}(G_{on,j}+P_{n,j})-L_{on}-T_{on}-E_n+10\left(\sum_{j=1}^{J_n}\beta_{1n,j}-\beta_{2n}-B_n\right)(F-F_o) = \frac{2S_{on}}{F_o}\frac{dF}{dt} \tag{23}$$

From Equation (23) we can see what the basic regulating dynamics in each of the control areas are. The primary regulating effects are supplemented by secondary regulation. Effective load and effective net interchange schedule will depend on frequency. Unit governor actions will seek to reduce $P_{n,j}$ to zero for all values of j, and supplementary control will adjust $G_{on,j}$ for all participating units to reduce E_n to zero. When these latter two conditions have been fulfilled, or if for any reason no further change can be made in an area to $\Sigma P_{n,j}$ or to E_n, then $\frac{dF}{dt}$ will be zero, and a steady state will have been achieved.

Supplementary controls will act more slowly than primary controls. In effect, their function may be regarded as reallocating generation changes to appropriate areas after random generation changes have been made by primary control, and restoring stored energy to achieve correction of system frequency after there has been transient primary support by stored energy change.

TIME DEVIATION AND INADVERTENT INTERCHANGE

When any area has E_n, ϕ_n, or τ_n errors, there will be corresponding deviations in system frequency, and deviations from net interchange schedules in all areas. The resultant system time deviation and area inadvertent interchange accumulations will be[5], respectively, as follows:

$$\varepsilon = -\frac{360}{F_r}\sum_{n=1}^{N}\frac{Y_n}{B_n}\left(\int_0^t E_n dt + \int_0^t \tau_n dt - 10B_n\int_0^t \phi_n dt\right) \tag{24}$$

$$I_n = \begin{cases} (1-Y_n)\left(\int_0^{t'} E_n dt+\int_0^{t'} \tau_n dt\right)-(1-Y_n)\left(10B_n\int_0^{t'} \phi_n dt\right) & \left.\right\} \text{Effects of Local Area Errors} \\ -Y_n \sum\limits_{\substack{i=1 \\ i\neq n}}^{N}\left(\int_0^{t'} E_i dt+\int_0^{t'} \tau_i dt\right)+10B_n \sum\limits_{\substack{i=1 \\ i\neq n}}^{N} Y_i \int_0^{t'} \phi_i dt & \left.\right\} \text{Effects of Remote Area Errors} \end{cases} \tag{25}$$

Time Deviation Correction

The accepted technique for coordinated system-wide correction of time deviation accumulations[5] is to add a time deviation bias term, common to all areas, to the area control error equation, thusly:

$$E_n = (T_n - T_{on}) - 10B_n(F - F_o - b\varepsilon) \tag{26}$$

The time deviation bias may be regarded as a further adaptive shift of the net interchange schedule as follows:

$$T_{yon} = T_{xon} + 10B_n(-b\varepsilon) \tag{27}$$

so that,

$$E_n = T_n - T_{yon} \tag{28}$$

Inadvertent Interchange Correction

A suggested technique for coordinated system-wide inadvertent interchange accumulation[5] is to add an inadvertent interchange term to the area control error equation in all areas, as follows:

$$E_n = (T_n - T_{on} + \frac{I_n}{H}) - 10B_n(F - F_o - b\varepsilon) \tag{29}$$

In Equation (29) there is a still further adaptive shift of the area net interchange schedule, the new schedule being given by:

$$T_{zon} = T_{yon} + \frac{I_n}{H} \tag{30}$$

so that,

$$E_n = T_n - T_{zon} \tag{31}$$

The three adaptive shifts of net interchange schedule, by frequency deviation, by time deviation and by inadvertent interchange accumulation are shown schematically in Figure 3, at (a), (b) and (c) respectively.

Magnitude of Frequency Bias

A brief comment might be made here concerning the magnitude of the frequency bias, B_n. Standard practice is to set B_n to be at least equal to β_n. The basis for this practice[6] has been the recognition that the compelling criterion for determining bias magnitude relates to assistance to be automatically provided to an area in need when there is a frequency deviation. Bias magnitude is relatively unimportant for other functions of net interchange bias control. A recent paper[7] suggested that optimum bias settings would be $B_n = \beta_n/2$. As noted in discussions in that reference, this was conventional practice prior to 1956, and was found at that time to result in undesirable results[8]. The recent recommendation apparently stems from a misconception of B_n as a regulating gain instead of an adaptive shift of schedule to assist remote areas.

Inspection of Equation (23) will reveal how, when $B_n = \beta_n$, governing response in area n to a remote load or generation change will persist as assistance to the remote area, as it should, until the remote area is able itself to fully respond to the change. See Reference 8 for additional discussion on the influence of bias settings.

ECONOMIC AND SECURITY DISPATCH

We have now accounted for all system objectives, and the first five area objectives that were tabulated in the introduction. In making generation allocations in the area to reduce E_n to zero, it will be desired to make the allocations in a manner that will simultaneously optimize area economy and security.

One technique for doing this is illustrated in Figure 4. Area control error is combined with total area generation to establish total generation required of the area. On-line computations are made, giving weight to telemetered data on prevailing system parameters and to stored data, and making optimum allocation of total needed area generation to available generating sources. Economy consistent with security constraints is the objective of the optimizing computations. Installations on order utilize advanced techniques of state estimation, sensitivity analysis and contingency studies[9] and security-oriented load flow studies[10] to compute optimized allocations.

CONTROL EXECUTION

Having established the desired generation level for each participating source as suggested by Figure 4, the next step is to adjust the speed changing motors of those sources so that actual generation will match desired generation. The preferred technique--identified as a "permissive" execution--is to let control action originate from E_n, using $G_{dn,j}$ vs $G_{n,j}$ comparisons to operate gates which will permit units to receive control signals that will at once act to reduce E_n to zero and establish $G_{n,j} = G_{dn,j}$ for all values of j.

Analog and digital systems, or combinations of the two, are used for such control execution[9]. Digital computer installations, with their extensive capabilities, can use advanced control techniques to achieve, in coordinated fashion, these dual control objectives[11,12,13].

REFERENCES

1. Definitions of terminology for automatic generation control on electric power systems, IEEE Publication, No. 94, July/August 1970.

2. Nathan Cohn, Fundamentals for control of generation and power flow on interconnected systems, presented at the Seminar on Control of Interconnected Systems, sponsored by the South American Comision de Integracion Electrica Regional, Lima, Peru, November 1970. L&N Reprint No. E7.0019 RP.

3. Nathan Cohn, Considerations in the regulation of interconnected areas, IEEE Transactions on Power Apparatus and Systems, PAS-86, No. 12, December 1967.

4. Nathan Cohn, Control of generation and power flow on interconnected systems, Wiley, 1966.

5. Nathan Cohn, Techniques for improving the control of bulk power transfers on interconnected systems, IEEE Transactions on Power Apparatus and Systems, PAS-90. No. 6, November/December 1971.

6. Nathan Cohn, Some aspects of tie-line bias control on interconnected power systems, Trans. Am. Inst. Elec. Engrs., 75, Pt. III, 1415-1428, 1956.

7. O. I. Elgerd and C. E. Fosha, Jr., Optimum megawatt-frequency control of multiarea electric energy systems, IEEE Transactions on Power Apparatus and Systems, Vol. PAS-89, No. 4, 556-563, April 1970.

8. Nathan Cohn, Bias revisited, presented to NAPSIC East Central Systems Group, Spring Meeting, St. Joseph, Michigan, April 1970. L&N Reprint No. E7.0014 RP.

9. Nathan Cohn, S. B. Biddle, Jr., R. G. Lex, Jr., E. H. Preston, C. W. Ross and D. R. Whitten, On-line computer applications in the electric power industry, IEEE Proceedings, Special Issue on Computers and Industrial Process Control, January 1970.

10. B. F. Wollenberg, The operator-oriented load flow and security dispatch, presented at the IEEE-PICA Conference, May 1969.

11. C. W. Ross, A comprehensive direct digital load-frequency controller, presented at the IEEE-PICA Conference, May 1967.

12. W. O. Stadlin, Economic allocation of regulating margin, IEEE Transactions on Power Apparatus and Systems, PAS-90, No. 6, November/December 1971.

13. C. W. Ross and T. A. Green, Dynamic performance evaluation of a computer controlled electric power system, presented at IEEE Power Engineering Society Summer Meeting and International Symposium on High Power Testing, Portland, Oregon, July 1971. To be published in IEEE Transactions.

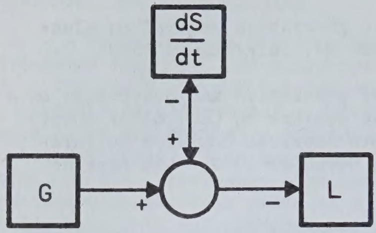


Figure I

System Generation-Load Balance

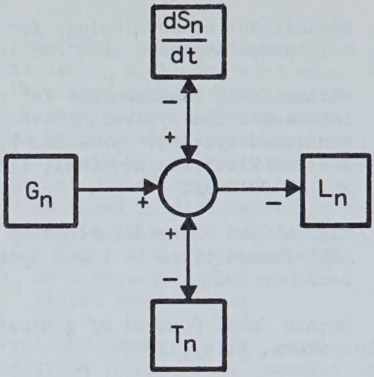

dSn/dt
−
+
Gn
+
−
Ln
+
−
Tn

Figure 2

Area Generation-Load Balance

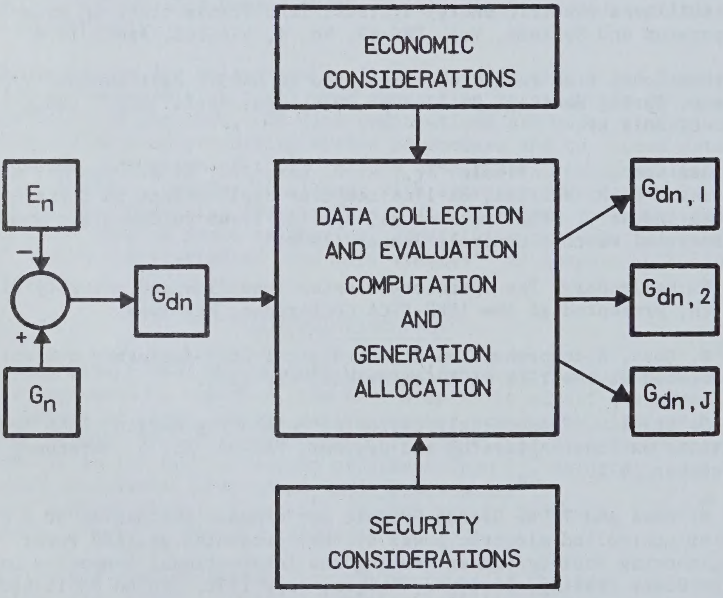

ECONOMIC CONSIDERATIONS
En
−
+
Gn
Gdn
DATA COLLECTION AND EVALUATION COMPUTATION AND GENERATION ALLOCATION
Gdn,1
Gdn,2
Gdn,J
SECURITY CONSIDERATIONS

Figure 4

Combining Regulation From Area Control Error
With Economic And Security Dispatch

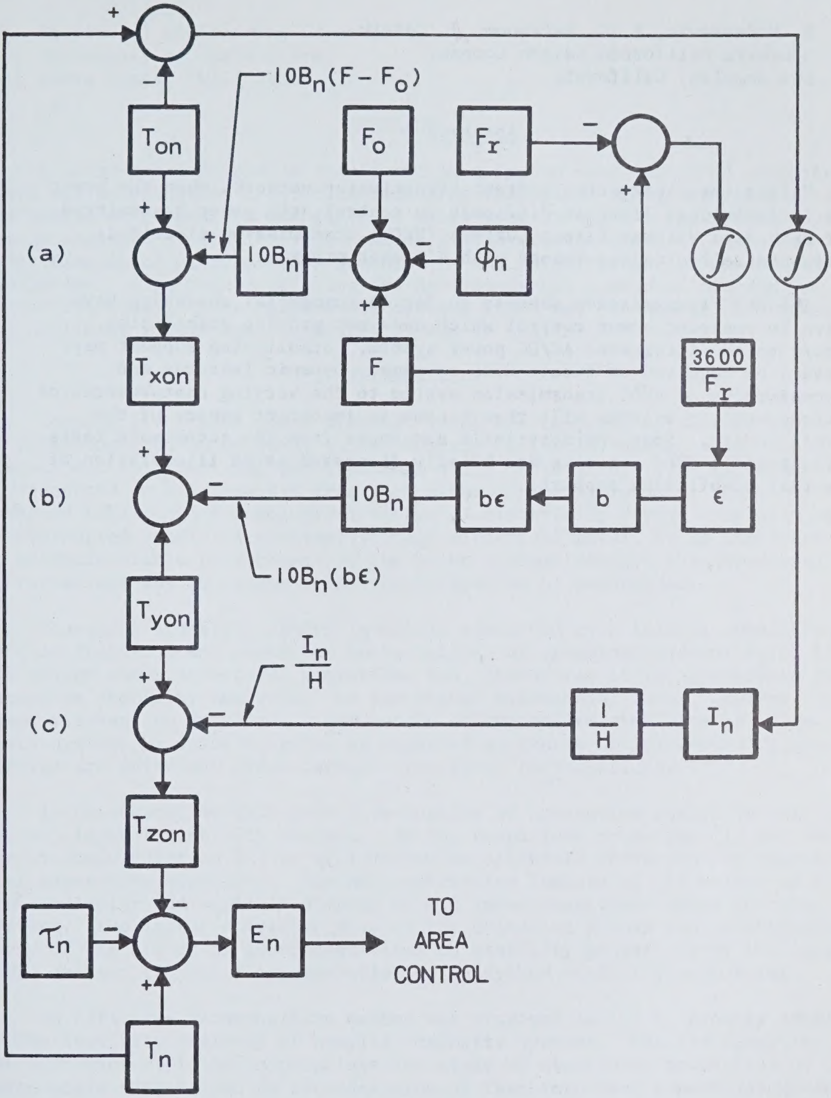

10Bn(F − Fo)
Ton
Fo
Fr
(a)
10Bn
ϕn
Txon
F
3600 / Fr
(b)
10Bn
bϵ
b
ϵ
10Bn(bϵ)
Tyon
In / H
(c)
1 / H
In
Tzon
τn
En
TO AREA CONTROL
Tn

Figure 3
Computation Of Area Control Error
With Adaptive Shifts Of
Area Net Interchange Schedule

HIGH VOLTAGE DIRECT CURRENT TRANSMISSION (HVDC) AND POWER SYSTEM STABILITY

V. Madzarevic, E. G. Pereboom, J. Sabath
Southern California Edison Company
Los Angeles, California

Abstract

Unlike the alternating current transmission network, when the power flow in individual lines is difficult to control, the power transmitted over each High Voltage Direct Current (HVDC) transmission circuit is always controlled in accordance with a preset program.

The HVDC transmission schemes so far in commercial operation have basically constant power control which does not provide stabilizing support to the integrated AC/DC power system. Stabilizing support may, however, be required on future HVDC systems. Dynamic features and responsiveness of HVDC transmission system to the varying disturbances of electric utility systems will then become an important aspect of the overall design. Some characteristic sequences from the acceptance tests on the Pacific HVDC Intertie are briefly discussed as an illustration of potential stabilizing support.

ON LARGE-SCALE SYSTEM STABILITY*

D. D. Šiljak
Electrical Engineering Department
University of Santa Clara
Santa Clara, California

ABSTRACT

A large-scale system is considered as a system constituted of subsystems which may be connected or disconnected from each other during operation. A new notion of *connective stability* is introduced by which a large-scale system is regarded as stable if it remains stable (in the sense of Liapunov) under *structural perturbations* produced by the on-off participation of the subsystems. Algebraic conditions are developed that guarantee exponential connective stability in large-scale systems which may be composed of linear and nonlinear sybsystems coupled by linear or nonlinear connections.

I. INTRODUCTION

Complex electrical power systems, which are composed of a large number of interconnected subsystems, quite commonly, either by design or by fault, are subject to structural perturbations which cause groups of subsystems to be connected or disconnected from each other during operation. Since numerous public and technological usages of electricity depend crucially on uninterrupted supply of constant voltage electrical power, it is important to maintain stable performance of the power systems despite the structural perturbations due to random on-off participation of subsystems.

Classical stability theory is mainly concerned with initial condition, forcing function, and parameter perturbations of dynamical systems which do not change their structural properties and, therefore, it is appropriate for subsystem stability analysis. On the higher hierarchial level, however, it seems relevant to introduce a new notion of connective stability of large-scale systems by which a system is regarded as stable if its stability properties are invariant under certain structural perturbations.

In this work, we will give a definition of connective stability relying on the Liapunov stability concept. By the comparison principle [1] and vector Liapunov function [2] we will derive an algebraic criterion for exponential connective stability. The most attractive feature of the criterion is its simplicity and explicit display of all interconnections among the subsystems. The latter characteristic of the criterion allows a straightforward study of the effect of interconnections on stability properties of the large-scale systems without reconsideration of subsystem stability properties.

An efficient decomposition method was proposed in [3] to greatly simplify the stability analysis of complex composite systems. The decomposition method, however, is not appropriate for study of structural properties of large-scale systems and no consideration of that important aspect was given. By avoiding some unnecessary complications in applying the comparison principle, we will improve in this work the decomposition method of [3] and make it

---

*The research reported herein was supported by NASA Grant NGR 05-017-010.

more suitable for studying stability of large-scale systems with invariant structure. Then, the improved technique can be used to determine stability of high-order subsystems which are parts of a large-scale system with changing structure.

II. SYSTEM DESCRIPTION

Let us consider a continuous dynamic system S described by the vector differential equation

$$\dot{x} = f(t, x) , \tag{S}$$

where $x(t) \in R^n$ is the state of the system, and the function $f: T \times R^n \to R^n$ is sufficiently smooth, that is, $f(t, x) \in C^{(0, 1)}$ on $T \times R^n$, so that the solutions $x(t; t_0, x_0)$ of (S) exist and are unique and continuous for all initial conditions $(t_0, x_0) \in T \times R^n$ and $t \in T_0$. T is the time interval $(\tau, +\infty)$, where τ is a number or the symbol $-\infty$, and T_0 is the semi-infinite time interval $[t_0, +\infty)$. Moreover, $f(t, 0) = 0$ for all $t \in T$ and the origin $x = 0$ of the state space R^n is the unique equilibrium point of (S).

The system S is considered as composed of s dynamic subsystems S_i described by vector differential equations

$$\begin{aligned} \dot{x}_i &= f_i(t, x_i) + \sum_{j=1}^{s} e_{ij} g_{ij}(y_j) \\ & \qquad\qquad\qquad\qquad (i = 1, 2, \dots, s). \\ y_i &= h_i(x_i) \end{aligned} \tag{S_i}$$

In (S_i), $x_i(t) \in R^{n_i}$ is the state of the i-th subsystem; $y_i(t) \in R^{m_i}$ is the output of the i-th subsystem; $f_i: T \times R^{n_i} \to R^{n_i}$; and the functions $g_{ij}: R^{m_j} \to R^{n_i}$ and $h_i: R^{n_i} \to R^{m_i}$ are such that

$$||g_{ij}[h_i(x_j)]|| \le \xi_{ij} ||x_j|| , \; (\forall\, i, j = 1, 2, \dots, s) . \tag{1}$$

where ξ_{ij} are nonnegative numbers, and $||x_j|| = (x_j^T x_j)^{1/2}$ denotes the Euclidean norm of the vector x_j. (For more general case see Appendix.)

Therefore, the state $x = (x_1^T\, x_2^T \dots x_s^T)^T$ of the system S is a column n-vector which contains all the states of the subsystems as partitions.

Accordingly, $n = \sum_{i=1}^{s} n_i$ is the dimension of S, and n_i is the dimension of S_i.

The second term in the first equation of (S_i) represents the actions of the s subsystems on the i-th subsystem. This term contains the *interconnection numbers* e_{ij} which can take on values zero or one with the meaning:

$$e_{ij} = \begin{cases} 1, & S_j \text{ acts on } S_i \text{ through } g_{ij} \\ 0, & S_j \text{ does not act on } S_i \end{cases} \tag{2}$$

The interconnection numbers e_{ij} may be viewed as binary elements of an $s \times s$ *interconnection matrix* E which displays the interconnections of subsystem S_i in the system S. Since groups of subsystems can be disconnected from each other and again connected together, the interconnection matrix changes its form. Structural changes that are of interest in this work are such that the interconnections between the groups are either totally "on" or totally "off", with the provision that if a single system constitutes a group, then its on-off feedback can be "on" or "off". It is important to note that the described structural changes may be arbitrary functions of the state $x(t) \in R^n$ and/or time $t \in T_0$.

To express symbolically the above defined structural perturbations, let us consider a real $s \times s$ matrix $A = (a_{ij})$ and specify its $q \times q$ principal submatrix A_q as

$$A_q = A\begin{pmatrix} i_1 i_2 \ldots i_q \\ i_1 i_2 \ldots i_q \end{pmatrix} = \begin{bmatrix} a_{i_1 i_1} & a_{i_1 i_2} & \ldots & a_{i_1 i_q} \\ a_{i_2 i_2} & a_{i_2 i_2} & \ldots & a_{i_2 i_q} \\ \ldots & \ldots & \ldots & \ldots \\ a_{i_q i_1} & a_{i_q i_2} & \ldots & a_{i_q i_q} \end{bmatrix}, \tag{3}$$

where $1 \leq i_1 < i_2 < \ldots < i_q \leq s$, and $1 \leq q \leq s$.

The determinant $\mu\begin{pmatrix} i_1 i_2 \ldots i_q \\ i_1 i_2 \ldots i_q \end{pmatrix}$ of the principal submatrix A_q is called the principal minor.

We introduce the notion of the *fundamental interconnection matrix* E_f of a given system S as the interconnection matrix in which all elements that correspond to existing connections between subsystems are set to one. Nonexisting interconnections are represented by "invariant zero elements". Then, any interconnection matrix, which specifies different connections between the subsystems, is obtained from E_f by replacing some of the ones by zeros. The "stiff" interconnections that are always "on" are represented by "invariant unit elements" and need no special considerations. However, it may appear useful to treat a group of subsystems with stiff interconnections as a single subsystem.

The structural changes that are considered in the following development give rise to interconnections which are represented by *principal interconnection matrices* E_p. These matrices are defined as interconnection matrices which can be transformed by only row and column transpositions into the quasidiagonal matrix

$$\begin{bmatrix} E_{q_1} & & & \\ & E_{q_2} & & \bigcirc \\ & & \ddots & \\ \bigcirc & & & E_{q_\ell} \end{bmatrix} \tag{4}$$

with nonzero principal submatrices $E_{q_1}, E_{q_2}, \dots, E_{q_\ell}$, and which are equal to the corresponding principal submatrices of the fundamental interconnection matrix E_f. Physically, matrix (3) shows that the system *S* is decomposed into (composed of) ℓ independent groups of subsystems S_i. The equality of the corresponding principal submatrices of E_p and E_f guarantees that the interconnections between the subsystems in each group remain the same as those represented by E_f.

To illustrate the structural changes that can take place, consider a system *S* given on Fig. 1. The system *S* has three subsystems S_1, S_2, S_3, and the corresponding fundamental matrix is

$$E_f = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{5}$$

Three possible interconnection matrices are

$$\underset{(a)}{\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \quad \underset{(b)}{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}} \quad \underset{(c)}{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}}. \tag{6}$$

(a) System *S* is decomposed into (composed of) two groups of subsystems, namely one formed by S_1 and S_2 with their interconnections as those in E_f, and the other formed by the single subsystem S_3 with its feedback loop closed. The principal submatrices $E\binom{12}{12}$ and $E\binom{3}{3}$ are equal to those of E_f in (5). The matrix (6a) *is* a principle interconnection matrix.

(b) All the interconnections are "off" and the system *S* is decomposed into three independent sybsystems S_1, S_2, S_3 with their on-off feedback loops open. There are no nonzero principle submatrices and the matrix (6b) *is* a principle interconnection matrix.

(c) The system *S* is decomposed into (composed of) two independent groups of systems as in (a) above. However, the interconnections in the group formed by S_2 and S_3 are not the same as represented by E_f in (5). Consequently, the nonzero principle submatrix $E\binom{23}{23}$ *is not* equal to that of E_f and the matrix (6c) *is not* a principle interconnection matrix.

III. CONNECTIVE STABILITY

It is of interest to incorporate the above defined structural changes in a stability analysis of the system S. We would like to have stability of S invariant under the described structural changes. Thus, we introduce:

Definition 1: *A free dynamic system S is connectively stable if and only if its equilibrium position is stable in the sense of Liapunov for all principle interconnection matrices.*

The above definition is broad enough to allow different kinds of Liapunov stability to be involved in a system analysis.

It is important to note immediately that a *necessary condition for connective stability* is that each subsystem S_i should possess the same stability properties as required from the system S. This follows from Definition 1 when the principal interconnection matrix is the zero matrix.

Of particular interest in this work is the following:

Definition 2: *A free dynamic system S is exponentially connectively stable if and only if there exist two positive numbers α and β independent of the initial conditions (t_0, x_0) such that*

$$||x(t;\, t_0,\, x_0)|| \leq \alpha\, ||x_0||\, e^{-\beta(t-t_0)}\,, \quad \forall t \in T_0\,, \tag{7}$$

for all $(t_0, x_0) \in T \times R^n$ and all principal interconnection matrices.

IV. ALGEBRAIC CRITERION

To derive conditions for exponential connective stability, we assume that each subsystem S_i is exponentially stable. Then, we consider a function $v_i\colon T \times R^{n_i} \to R^1$ such that $v_i(t, x_i) \in C^{(1,1)}$ on $T \times R^{n_i}$, and make use of the following:

Theorem 1: *A free dynamic subsystem $S_i(e_{ij} = 0\,,\ \forall\, j = 1, 2, \ldots, s)$ is exponentially stable if and only if there exists a positive definite function $v_i(t, x_i)$ on $T \times R^{n_i}$ such that*

$$\begin{aligned} &\eta_{i1}\, ||x_i|| \leq v_i \leq \eta_{i2}\, ||x_i||\,, \\ &v_i \leq -\,\eta_{i3}\, ||x_i||\,, \\ &||\ \mathrm{grad}\ v_i\ || \leq \eta_{i4}\,, \end{aligned} \tag{8}$$

where $\eta_{i1}, \eta_{i2}, \eta_{i3}, \eta_{i4}$ *are positive numbers, and*

$$\dot{v}_i = \frac{\partial}{\partial t} v_i + (\text{grad } v_i)^T f_i \quad . \tag{9}$$

Theorem 1 follows directly from a result of Krasovskii [4] when one considers $v_i^{1/2}$ instead of v_i . This modification, however, enables us to represent faithfully the actual physical interconnections among the subsystems and solve effectively the problem of connective stability.

Algebraic criterion for exponential connective stability is now given by:

Theorem 2: A free dynamic system S is exponentially connectively stable if the elements of the real $s \times s$ matrix $A = (a_{ij})$ with elements

$$a_{ij} = -\delta_{ij}\eta_{i2}^{-1}\eta_{i3} + e_{ij}\xi_{ij}\eta_{j1}^{-1}\eta_{i4} \; , \tag{10}$$

satisfy the inequalities

$$(-1)^q \, \mu\begin{pmatrix}12\ldots q\\ 12\ldots q\end{pmatrix} > 0 \; , \; (\forall q = 1, 2, \ldots , s) \tag{11}$$

for the fundamental interconnection matrix E_f .

In (8), δ_{ij} is the Kronecker delta. According to (3), inequalities (11) are

$$a_{11} < 0 \; , \; \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} > 0, \; \ldots \; , \; (-1)^s \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1s} \\ a_{21} & a_{22} & \cdots & a_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ a_{s1} & a_{s2} & \cdots & a_{ss} \end{bmatrix} > 0 \; . \tag{12}$$

To prove Theorem 2, we need first to calculate the total time derivative $\dot{\underset{\sim}{v}}_i$ of the function v_i along the solutions of (S_i) ,

$$\dot{\underset{\sim}{v}}_i = \dot{v}_i + (\text{grad } v_i)^T \sum_{j=1}^{s} e_{ij} g_{ij}[h_j(x_j)] \; . \tag{13}$$

By applying inequalities (1) and (8), we can rewrite (12) as

$$\dot{\underset{\sim}{v}}_i \leq -\eta_{i2}^{-1}\eta_{i3} v_i + \eta_{i4} \sum_{j=1}^{s} e_{ij}\xi_{ij}\eta_{j1}^{-1} v_j \; ; \tag{14}$$

and each interconnection g_{ij} of a subsystem S_i is related to the corresponding function v_j .

Now, we define an s-vector v as $v = (v_1 v_2 \ldots v_s)^T$, and using (14) form the differential inequality

$$\dot{v} \leq Av \ , \tag{15}$$

where the elements a_{ij} of the matrix A are given by (10) of Theorem 2. The vector inequality (15) is to be interpreted as component-by-component majorization. (That is, if $a, b \in R^s$ and $a \geq b$ then $a_i \geq b_i$ for all $i = 1, 2, \ldots, s$).

It is important to note that each nondiagonal element $a_{ij} (i \neq j)$ of the matrix A in (15) is related to a single interconnection between the corresponding subsystems S_i and S_j, a feature that is not present in the matrix A of Bailey [3]. This particular characteristic of the matrix A here, which is apparent from expression (14), makes it possible to solve the connective stability problem in large-scale systems.

We now need the well-known result [2]:

Lemma 1: *Let* $v(t; t_0, v_0)$ *be a solution of the differential inequality* (15) *for* $v_0 = v(t_0; t_0, v_0)$, *and let* $r(t; t_0, r_0)$ *be a solution of the comparison equation*

$$\dot{r} = Ar \tag{16}$$

for $r_0 = r(t_0; t_0, r_0)$. *If* $v_0 = r_0$, *and the elements of the real* $s \times s$ *matrix* $A = (a_{ij})$ *are*

$$a_{ij} = \begin{cases} < 0 \ , & i = j \\ \geq 0 \ , & i \neq j \end{cases} \qquad (\forall i \ , \ j = 1, 2, \ldots, s) \ , \tag{17}$$

then,

$$v(t; t_0, v_0) \leq r(t; t_0, v_0) \ , \ \forall t \in T_0 \ . \tag{18}$$

To determine stability of the linear comparison equation (16), we use the following result of Sevastyanov [5]:

Lemma 2: *A real* $s \times s$ *matrix* $A = (a_{ij})$ *with* $a_{ij} \geq 0$, $(\forall i \ , \ j = 1, 2, \ldots, s; \ i \neq j)$ *has all eigenvalues* λ_k *with negative real parts,*

$$\text{Re}\ \lambda_k < 0 \ , \ (\forall k = 1, 2, \ldots, s) \ , \tag{19}$$

if and only if the inequalities (11) *are satisfied.*

From Lemma 1 and 2, it follows that the satisfaction of algebraic conditions (11) guarantees the asymptotic stability of the system S for the fundamental interconnection matrix E_f. It is left to show that the stability is also exponential and independent of the form of principal interconnection matrices. The latter means that exponential stability for the fundamental matrix implies exponential stability for any principal interconnection matrix - this is exponential connective stability.

Let

$$\sigma = \max_k \operatorname{Re} \lambda_k \, , \, (k = 1, 2, \dots , s) \, . \tag{20}$$

Then [6], there exist two positive numbers $\varepsilon < |\sigma|$ and $\rho = \rho(\varepsilon)$ such that

$$||e^{A(t-t_0)}|| \leq \rho \, e^{(\sigma+\varepsilon)(t-t_0)} \quad , \quad \forall t \in T_0 \, , \tag{21}$$

where the Euclidean norm is used, $||e^{A(t-t_0)}|| = \{tr[e^{A(t-t_0)}]^T[e^{A(t-t_0)}]\}^{1/2}$. By applying inequalities (8), (21), and various technicalities involving Euclidean norms, one can show that from (19) and (20) follows

$$||x(t; t_0, x_0)|| \leq \alpha ||x_0|| e^{-\beta(t-t_0)} \, , \, \forall t \in T_0 \tag{22}$$

for all $(t_0, x_0) \in T \times R^n$, where

$$\alpha = \rho \eta_1^{-1} \eta_2, \ \beta = - \sigma - \varepsilon \, , \tag{23}$$

and

$$\eta_1 = \min_i \eta_{i1} \, , \, \eta_2 = \max_i \eta_{i2} . \tag{24}$$

This means that the satisfaction of inequalities (11) assures the exponential stability of the system S .

Finally, we observe an important property of inequalities (11) expressed by:

Lemma 3: *If an* $s \times s$ *matrix* $A = (a_{ij})$ *with* $a_{ij} \geq 0$, $(\forall i, j = 1, 2, \dots , s; i \neq j)$ *satisfies the inequalities* (11), *then each principal submatrix* $A\begin{pmatrix} i_1 i_2 \dots i_q \\ i_1 i_2 \dots i_q \end{pmatrix}$ *has all its eigenvalues with negative real parts.*

By adjusting the signs in Kotelyanskii's generalization [7] of the Sevastyanov result given in Lemma 2, we can derive the following:

Lemma 4: *If in a real* $s \times s$ *matrix* $A = (a_{ij})$ *with* $a_{ij} \geq 0$, $(\forall i, j = 1, 2, \dots , s; i \neq j)$ *successive principal minors satisfy condition* (11), *then all principle minors satisfy the condition*

$$(-1)^q \mu\begin{pmatrix} i_1 i_2 \dots i_q \\ i_1 i_2 \dots i_q \end{pmatrix} > 0 \, , \, (1 \leq i_1 < i_2 < \dots < i_q \leq s \, , \ \forall q = 1, 2, \dots , s). \tag{25}$$

From Lemma 4 follows Lemma 3 and, therefore Theorem 2. This completes the proof of Theorem 2.

V. CONCLUSION

A useful aspect of the solution to the connective stability problem of large-scale systems with changing structure, which is offered by Theorem 2, is the algebraic criterion. Due to the nonnegativity property of the comparison stability matrix A, the criterion leads to a simpler stability test than the Routh-Hurwitz or Liapunov matrix test. Certainly, the most attractive characteristic of the proposed stability criterion is the simplicity of the comparison stability matrix A in which each element a_{ij} represents a single interconnection between the subsystems.

Exponential stability of each subsystem is a necessary condition for exponential connective stability of the corresponding large-scale system. This is not too restrictive, since this kind of stability for subsystems is usually the most desirable stability property in practical situations. Moreover, large classes of linear and nonlinear system are exponentially stable [3,8,9], or can be designed to have this characteristic [10].

The obtained results can be used advantageously also in the situations when a large-scale system is decomposed in subsystems to simplify the stability analysis as proposed by Bailey [3]. One needs to consider only the fundamental interconnection matrix and use the proposed algebraic test to conclude stability of the large-scale system. The results proposed here are also less conservative than those obtained by Bailey [3], since less number of majorizations is used.

Further extensions of this theory are currently performed to include connective stability of discrete dynamic systems, connective absolute stability, and some aspects of finite regions of connective stability.

Applications of the theory are conducted in the areas of power systems and altitude control systems for space vehicles.

ACKNOWLEDGEMENT

The author is indebted to Professor Lj. Grujić, Mechanical Engineering Department, University of Belgrade, Yugoslavia, for suggesting Theorem 1, and for many useful discussions of this work. The author is also grateful to Professor S. Weissenberger, Mechanical Engineering Department, University of Santa Clara, Santa Clara, California, for his valuable comments on this paper.

REFERENCES

1. Conti, R., "Sulla prolungbilità delle soluzioni di un sistema di equazioni differenziali ordinarie", Boll. Unione Mat. Ital., Vol. 11, 1956, pp. 510-514.

2. Bellman, R., "Vector Lyapunov Functions", J. SIAM Control, Ser. A, Vol. 1, No. 1, 1962, pp. 32-34.

3. Bailey, F. N., "The Application of Lyapunov's Second Method to Interconnected Systems", J. SIAM Control, Ser. A, Vol. 3, No. 3, 1966, pp. 443-462.

4. Krasovskii, N. N., "Some Problems of the Theory of Stability of Motion" (in Russian), FIZMATGIZ, Moscow, 1959. (Translation: "Stability of Motion", Stanford University Press, Stanford, California, 1963.)

5. Sevastyanov, B. A., "Theory of Branching Stochastic Processes" (in Russian), Uspehi Mat. Nauk, Vol. 6, No. 6, 1951, pp. 47-99.

6. Demidovich, B. P., "Lectures on Mathematical Stability Theory"(in Russian), Nauka, Moscow, 1967.

7. Koltelyanskii, I. N., "On Some Properties of Matrices with Positive Elements" (in Russian), Mat. Sb., Vol. 31, No. 3, 1952, pp. 497-506.

8. Šiljak, D. D., "Nonlinear Systems", Wiley, 1969.

9. Šiljak, D. D. and Sun, C. K., "Exponential Absolute Stability of Discrete Systems", ZAMM, Vol. 51, No. 4, July, 1971, pp. 271-275.

10. Karmarkar, J. S. and Šiljak, D. D., "A Computer-Aided Regulator Design", presented at this conference.

APPENDIX

It is of interest to point out that the obtained results are applicable to the more general case of interconnections between subsystems S_i when S_i is described by

$$\begin{aligned} \dot{x}_i &= f_i(t, x_i) + g_i(t, e_{i1}y_1, e_{i2}y_2, \dots, e_{is}y_s) \\ & \qquad\qquad (i = 1, 2, \dots, s), \qquad (S_i') \\ y_i &= h_i(t, x_i) \end{aligned}$$

where

$$||g_i(t, e_{i1}y_1, e_{i2}y_2, \dots, e_{is}y_s)|| \le \sum_{j=1}^{s} e_{ij}\xi_{ij}||x_j||, \ \forall t \in T$$

$$(\forall\, i = 1, 2, \dots, s). \qquad (1')$$

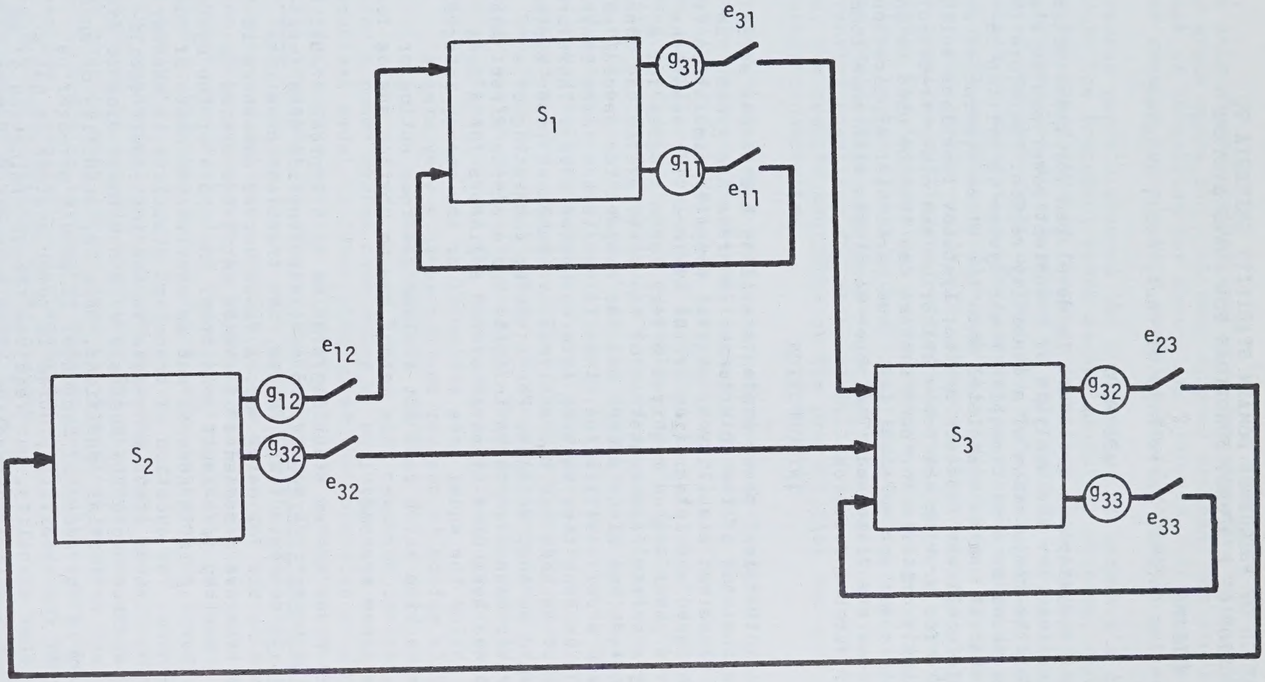

e_{31}
S_1
g_{31}
g_{11}
e_{11}
e_{12}
S_2
g_{12}
g_{32}
e_{32}
e_{23}
S_3
g_{32}
g_{33}
e_{33}

Fig. 1

THE APPLICATION OF FREQUENCY DOMAIN STABILITY CRITERIA TO THE CONSTRUCTION OF LYAPUNOV FUNCTIONS FOR POWER SYSTEMS

J. L. WILLEMS
Engineering School, University of Gent, Gent, Belgium

ABSTRACT

In this paper a systematic procedure is developed for constructing Lyapunov functions for the analysis of transient power system stability and for the computation of a stability region. Therefore the mathematical model of a multimachine power system is written in such a way that it can be associated with the block diagram of a multivariable nonlinear feedback system; Lyapunov functions which are available for proving the Popov criterion and other frequency domain stability criteria for such systems can then be used to analyse transient power system stability. Some particular applications of of this idea are discussed. The paper concludes with some comment on areas for further research.

INTRODUCTION

The problem of transient power system stability has become more important as the size of the interconnected areas has grown. The analysis of transient stability by digital computer simulation (the routine technique) and integration of the differential equations describing the power system requires a very large program for a sufficiently accurate representation of the power system and considerable computer time since swings must be computed for several seconds to obtain a full oscillation; this is due to the large inertias involved and the long ties between interconnected areas. Therefore accurate direct methods for the analysis of transient power system stability would be most welcome. For a system consisting of a single generator connected to an infinite bus a simple direct analysis method has been used by power system engineers for a long time; it is called the equal area criterion or the energy integral criterion. This method is however restricted to a very simple model of the machine which does not include governor action or voltage regulation. Moreover this method has no obvious analog for multimachine power systems.

The transient power system stability problem is a typical stability problem in the sense of Lyapunov, in contradistinction with input-output stability concepts (10). Indeed, the transient stability problem arises in the following way. A fault occurs somewhere in the system; this triggers a sequence of events until the system is restored to a healthy post-fault condition. The state of the system after fault clearing is in general not an equilibrium state of the post-fault system. The question of transient stability is whether or not the power system state converges to the post-fault equilibrium; this requires hence the analysis of a nonlinear dynamic system with respect to initial conditions, this is, stability or asymptotic stability in the sense of Lyapunov. It should moreover be emphasized that the stability problem in power systems is not a problem of global stability, but requires the determination of an estimate of the domain of attraction of an equilibrium solution. Power systems are never asymptotically stable in the large. The second or direct method of Lyapunov (10) seems at the present time to be the only general method available for computing stability regions in the state space for nonlinear dynamic systems. The study and development of techniques for generating Lyapunov functions for power systems remains hence a very useful area of research.

The present paper discusses the application of Lyapunov's direct method to the transient power system stability problem, and in particular the generation of Lyapunov functions. For the application to power systems the direct method is only used to compute the stability region of the post-fault system; the post-fault initial state is determined by simulation and integration of the system equations during fault conditions. Although this method hence does not exclude computer simulation completely, the most time-consuming part of it (which requires cutting-and-trying) is replaced by a direct analysis technique. The introduction of the events occurring during fault conditions in the direct method (such that the initial state becomes the pre-fault state) would lead to the consideration of nonautonomous differential equations for which stability theory is much less developed; the results available at the present time for such systems are rather conservative.

This paper discusses a particular approach for generating a useful Lyapunov function. We attempt to apply stability theory developed for nonlinear multivariable feedback control systems to the power system stability problem and to reduce in this way the guess-work involved in constructing a suitable Lyapunov function. Therefore a multimachine power system is modeled as a multivariable control system with a linear element in the forward path and multiple uncoupled memoryless nonlinearities in the feedback path. The generalized Popov criterion and its available proofs by means of Lyapuov's second method, and also other frequency domain stability criteria, can then be used to generate a useful Lyapunov function for the power system stability problem. Therefrom an estimate of the domain of attraction of the equilibrium state under consideration is derived.

MATHEMATICAL MODEL

Consider a power system consisting of n synchronous machines. The mathematical model is formulated under the following assumptions :
(i) A synchronous machine is represented by a constant voltage behind its transient reactance; in other words it is assumed that rotor flux linkages remain constant during the transient phenomenon. Flux decay and voltage regulation are hence not taken into consideration. This assumption is valid in most practical cases since the time constants involved in these electromagnetic phenomena are large compared to the time period which is of interest in transient stability analysis.
(ii) The mechanical power input from the prime mover to any synchronous machine is assumed constant. The governor time constants are indeed usually much larger than the transient periods.
(iii) The transfer impedances between the machines are assumed to be reactive; transfer conductances are neglected. They are usually negligible for practical power systems.
(iv) The damping power is assumed to be proportional to the slip velocity; it is mainly due to asynchronous torques and mechanical friction; very fast governors also introduce similar damping.
To keep the mathematical expressions simple only cylindrical rotor

machines are considered in this paper. Then the motion of the i^{th} machine of the power system is governed by the second order differential equation

$$M_i d^2\delta_i/dt^2 + a_i d\delta_i/dt + P_{ei}(t) - P_{mi} = 0 \qquad (1)$$

where t denotes time, δ_i is the angle (in electrical degrees) between the rotor shaft and a shaft running at synchronous speed, M_i the inertia constant of the i^{th} machine, a_i the damping constant, P_{ei} the electrical power generated by the machine and P_{mi} the mechanical power input from the prime mover (which is assumed constant). The electrical power output equals

$$P_{ei} = G_i E_i^{\,2} + \sum_{j=1}^{n} E_i E_j Y_{ij} \sin(\delta_i - \delta_j) \qquad (2)$$

where E_i is the internal voltage of the i^{th} machine, G_i its short-circuit conductance, Y_{ij} the transfer admittance (assumed inductive) between the ith and the j^{th} machines.

The equilibrium states of the power system equations are the solutions of the set of 2n equations :

$$\omega_i = 0 \;, \; P_{ei} = P_{mi} \qquad (i = 1,...n) \qquad (3)$$

where $\omega_i = d\delta_i/dt$ is the difference between the actual speed and the synchronous speed. Although these equations only involve (2n-1) unknown elements (the speeds and the load angle differences), the set is not overdetermined. Indeed, one of the equations can be dropped, since a necessary condition for the existence of an equilibrium is that the total power input equals the power dissipated :

$$\sum_{i=1}^{n}(P_{mi} - E_i^{\,2} G_i) = 0$$

Hence of the equations of the set (3) is hence redundant. The power system is hence of order 2n-1, and not of order 2n; it can even be reudced to a system of order (2n-2) if it is undamped or uniformly damped (the definition of uniform damping is given below).

The set of equations describing the power system is transformed by simple substitutions into the canonical form of a vector differential equation of the first order having the same form as the equations governing a nonlinear feedback system. We obtain :

$$\dot{\underline{x}}(t) = \underline{A}\underline{x}(t) - \underline{B}\underline{f}(\underline{C}\underline{x}(t)) \qquad (4)$$

where the underlined symbols denote vectors or matrices, and where the following notations are introduced :

$$\underline{x} = \begin{bmatrix}\underline{\omega} \\ \underline{y}\end{bmatrix} \;, \; (\underline{\omega})_i = \omega_i \,, \quad (\underline{y})_i = y_i = \delta_i - \delta_i^o$$

δ_i^o is the load angle of the equilibrium state under consideration, and $\underline{x}$ is the state vector,

$$\underline{A} = \begin{bmatrix} \underline{M}^{-1}\underline{R} & \underline{O}_{nn} \\ \underline{I}_{nn} & \underline{O}_{nn} \end{bmatrix}, \quad \underline{C} = \begin{bmatrix} O_{mn} & \underline{D} \end{bmatrix}, \quad \underline{B} = \begin{bmatrix} \underline{M}^{-1}\underline{D}^{T} \\ \underline{O}_{nm} \end{bmatrix},$$

$$\underline{M} = \mathrm{diag}(M_1, M_2, \ldots, M_n), \quad \underline{R} = \mathrm{diag}(a_1, a_2, \ldots, a_n),$$

where $m = n(n-1)/2$, $\underline{I}_{nn}$ denotes the $(n \times n)$ identity matrix, $\underline{O}_{jk}$ the $(j \times k)$ zero matrix, $\underline{D}$ is an $(n \times m)$ matrix such that $\underline{z} = \underline{Dy}$ has the elements $y_1-y_2, y_1-y_3, \ldots, y_1-y_n, y_2-y_3, \ldots, y_{n-1}-y_n$, and T denotes transposition. $\underline{A}$ is a $(2n \times 2n)$, B a $(2n \times m)$, and C an $(m \times 2n)$ matrix. The vector function $\underline{f}(\underline{z})$ has m components and is of the diagonal or uncoupled type, which means that the i^{th} component of the vector function only depends on the i^{th} component of the argument $\underline{z}$; it is defined as follows

$$(\underline{f}(\underline{z}))_i = f_i(z_i) = E_p E_q Y_{pq} (\sin(z_i + z_i^o) - \sin(z_i^o))$$

where p and q are the indices of the components y_i on which z_i is dependent, and where $\underline{z}^o$ is the value of $\underline{z}$ for $y_i = \delta_i^o$.

Equation (4) is a familiar one in stability theory of nonlinear control systems. Indeed it describes the multivariable feedback system of Fig. 1 with a linear time-invariant element with transfer function matrix $\underline{G}(s) = \underline{C}(\underline{I}s-\underline{A})^{-1}\underline{B}$ in the forward path and m uncoupled time-invariant memoryless nonlinear elements in the feedback path. Numerous results have been reported on the stability properties of this type of control system. One of the outstanding results is the Popov criterion which yields sufficient conditions for the asymptotic stability in the large of the null solution of system (4). For power systems we never have asymptotic stability in the large; we are concerned with the computation of the domain of attraction of the null solution. Nevertheless the available techniques for proving the Popov criterion (and other frequency domain stability criteria) by means of a Lyapunov function can be used here to generate a useful Lyapunov function and to derive an estimate of the stability region in the state space.

GENERATION OF LYAPUNOV FUNCTIONS

An elegant procedure for generating a Lyapunov function for proving the Popov criterion for system (4) was outlined by Anderson (1). He proves that

$$V(\underline{x}) = \underline{x}^T\underline{Px} + 2 \int_{\underline{0}}^{\underline{Cx}} \underline{f}(\underline{z})^T\underline{Q}d\underline{z} \tag{5}$$

is a useful Lyapunov function for the null solution of (4), if $\underline{N}$ and $\underline{Q}$ are positive semi-definite matrices such that $(\underline{N}+\underline{Q}s)\underline{G}(s)$ is a positive real matrix, where $\underline{N}$ and $\underline{Q}$ are diagonal, and if $\underline{P}$ is a positive definite solution of the equations

$$\begin{aligned} \underline{PA} + \underline{A}^T\underline{P} &= -\underline{LL}^T \\ \underline{PB} &= \underline{C}^T\underline{N} - \underline{LW} + \underline{A}^T\underline{C}^T\underline{Q} \\ \underline{W}^T\underline{W} &= \underline{QCB} + \underline{B}^T\underline{C}^T\underline{Q} \end{aligned} \tag{6}$$

where $\underline{L}$ and $\underline{W}$ are auxiliary matrices. For the mathematical model obtained above the equations (6) simplify considerably since

$$\underline{CB} = \underline{B}^T\underline{C}^T = \underline{0}_{mm}$$

such that $\underline{W}$ also vanishes; the third matrix equation can be left out and the other two are uncoupled. The frequency domain stability condition of the Popov type is satisfied for $\underline{N} = \underline{0}$ and $\underline{Q} = \underline{I}$. We obtain the following Lyapunov functions

$$\text{(i)}\quad V_o(\underline{x}) = \sum_{i=1}^{n} M_i\omega_i^2/2 + \int_{\underline{0}}^{\underline{Cx}} \underline{f}(\underline{z})^T d\underline{z} \tag{7}$$

which is valid as a Lyapunov function for all (nonnegative) values of the damping constants;

$$\text{(ii)}\quad V_1(\underline{x}) = \sum_{i=1}^{n}\ \sum_{j=i+1}^{n} M_iM_j(\omega_i-\omega_j)^2/2(\sum_i M_i) + \int_{\underline{0}}^{\underline{Cx}} \underline{f}(\underline{z})^T d\underline{z} \tag{8}$$

which is only a valid Lyapunov function for uniform damping, that is, if

$$a_1/M_1 = a_2/M_2 = \ldots = a_n/M_n \;.$$

$$\text{(iii)}\, V_k(\underline{x}) = (1-k)V_o(\underline{x}) + kV_1(\underline{x}) \tag{9}$$

which is a Lyapunov function with negative semi-definite derivative if k is smaller than the positive root of the quadratic equation :

$$pk^2 + k - 1 = 0 \tag{10}$$

where p is the constant

$$p = \sum_{i=1}^{n}\ \sum_{j=i+1}^{n} (M_j\sqrt{a_i/a_j} - M_i\sqrt{a_j/a_i}\,)^2 \,/(2\sum_{i=1}^{n} M_i\,)^2$$

which vanishes for a uniformly damped system. It is easily checked that $k \leq 1$. It will be useful in the next section to note the equality

$$V_1(\underline{x}) = V_o(\underline{x}) - (\sum_{i=1}^{n} M_i\omega_i)^2 \,/2(\sum_{i=1}^{n} M_i)$$

and to remark that $V_k(\underline{x})$ reduces to $V_o(\underline{x})$ and $V_1(\underline{x})$ for k = 0 and 1 respectively.

TRANSIENT STABILITY REGION

From the Lyapunov function generated in the previous section the stability region bounded by

$$V(\underline{x}) = V(\underline{x}^u) \tag{11}$$

is obtained, where $\underline{x}^u$ is the unstable equilibrium state closest to the null solution, that is, corresponding to the smallest Lyapunov surface (11). It is easily seen that the stability region obtained increases with the value of the constant k, and is hence largest in the case of uniform damping. Note indeed

$$V_o(\underline{x}^u) = V_1(\underline{x}^u) \pm V_k(\underline{x}^u)$$

for all k. For the above determination of the stability region the fact is used that the set of equations (3) determining the equilibrium states is equivalent to the set

$$\partial V_k(\underline{x})/\partial\omega_i = 0\ ,\quad \partial V_k(\underline{x})/\partial\delta_i = 0$$

for all k.

The stability region obtained above is hence largest for uniform damping; the Lyapunov function (7) yields the most conservative estimate of the domain of attraction of the equilibrium state at the origin of the state space. This conclusion cannot be carried over to the true domains of attraction, since Lyapunov theory in general only leads to sufficient stability conditions. To study the effect of uniform damping on the domain of attraction, the following two-machine system was simulated on a digital computer :

$$d\omega_1/dt = -a_1\omega_1 + 0.5 - \sin(x+\pi/6)$$
$$d\omega_2/dt = -a_2\omega_2 - 0.5 - \sin(x+\pi/6)$$
$$dx/dt = \omega_1 - \omega_2$$

where x is the deviation of the load angle difference from its equilibrium value ($\pi/6$). The initial states $x(0) = 0$, $\omega_1(0) = \omega_2(0) = \omega_o$ are considered for various values of ω_o. For a uniformly damped system ($a_1=a_2=a$), the solution is easily seen to be

$$x(t) = 0,\ \omega_1(t) = \omega_2(t) = \omega_o e^{-at}.$$

This shows that the motions are converging for all positive a, and that the initial conditions under consideration belong to the domain of attraction of the equilibrium state at the origin. The case of non-uniform damping ($a_1=0.1$, $a_2=2.0$) is also simulated for the same set of initial conditions; convergent solutions are obtained for ω_o up to 1.5, but divergent solutions for $\omega_o = 1.75$ and higher. This proves that the generally accepted intuitive statement that damping always aids stability is not true in all cases. Results in the same direction are reported by McLain (4). However it remains to be investigated whether or not this phenomenon can occur in practical power systems.

Lyapunov function (7) and the stability region obtained from it are equivalent to the results of El-Abiad and Nagappan (3); Lyapunov function (8) was used by Ribbens-Pavella (5,6), and corresponds to the energy function considered by Aylett (2). The latter Lyapunov function yields a larger stability region than the former but it is only valid for uniformly damped systems. However the improvement obtained by using V_1 instead of V_o for uniformly damped systems is not as large as one might expect in practical applications. Indeed, consider such a system without transfer conductances; adding the n equations (1), we have

$$(d^2/dt^2 + c\ d/dt)\sum_i M_i\omega_i = 0$$

where $c = a_i/M_i$. Starting from a prefault equilibrium state ($\omega_i = 0$), it follows that $\Sigma M_i\omega_i$ vanishes along the solution such that Lyapunov functions (7) and (8) lead to the same stability condition and the same estimate of the critical switching time.

Simulation studies by Ribbens-Pavella (5,6) on a nine machine system show that the above obtained Lyapunov functions do not give conservative stability conditions, but yield very sharp estimates of the critical switching time. Note also that the stable equilibrium under consideration does not have to be computed accurately; one indeed easily checks that (11) does not depend on it. Further results can be obtained by using other $\underline{N}$ and $\underline{Q}$ matrices and by using other frequency domain stability criteria for nonlinear feedback systems.

FURTHER RESEARCH

Further research in the area of the application of Lyapunov theory to the transient power system stability problem is needed at least in the following two directions. First, the mathematical model should be improved (the above assumptions should be relaxed) to approach the real physical system more closely, and second, methods for the generation of Lyapunov functions should be developed to go along with improvements in the mathematical model.

It can easily be shown (8,11) that the above technique can easily be adapted to take linear governors into consideration. However not the same is true for flux decay, saturation, voltage regulation, and excitation dynamics. The difficulties stem from the fact that one obtains multiplicative type nonlinearities; for such systems the known stability criteria (e.g. the circle criterion) are much more conservative that the Popov criterion. It should also be noted that the direct method discussed in this paper requires the integration of the system equations during the fault period where a series of successive events occurs, such as circuit reclosing, generator and load dropping, and capacitor switching. If one would try to include the fault period in the direct stability analysis (such that the initial state is the prefault power system state), one obtains a nonautonomous nonlinear system for which stability theory is much less deveoped than for autonomous systems. Another aspect that requires further examination is better representation of the dynamics of the loads. Ribbens-Pavella (5) developed an approximate Lyapunov function for dealing with transfer conductances, and showed by simulations that the obtained stability conditions are good approximations.

CONCLUSION

In this paper it has been shown that under some assumptions a power system can be associated with a multivariable nonlinear feedback system, and that Lyapunov functions can be generated by means of existing frequency domain stability criteria for such systems. Although some interesting results have already been obtained on the application of Lyapunov's direct method to the transient power system stability problem, much research remains to be done to make this method competitive with computer simulation stability analysis.

REFERENCES

1. B.D.O. Anderson, "Stability of control systems with multiple nonlinearities," Journal of the Franklin Institute, vol. 282, 1966, pp. 156-160.

2. P.D. Aylett, "The energy-integral criterion of transient stability limits of power systems," Proc. Inst. Elec. Eng., vol. 105 (C), 1958, pp. 527-536.

3. A.H. El-Abiad and K. Nagappan, "Transient stability regions of multimachine power systems," IEEE Trans. Power App. Syst., vol. PAS-85, Feb. 1966, pp. 169-179.

4. D.K. McLain, discussion of J.L. Willems, "Optimum Lyapunov functions and transient stability regions for multimachine power systems," Proc. Inst. Elec. Eng., vol. 118, 1971, p.603.

5. M. Ribbens-Pavella, "Théorie générale de la stabilité transitoire de n machines synchrones," Ph. D. dissertation, University of Liège, Liège, Belgium, Oct. 1969.

6. ———, "Transient stability of multimachine power systems by Liapunov's direct method," IEEE Winter Power Meeting, Paper 71 CP 17 - PWR, Feb. 1971.

7. W. Siddiqee, "Transient stability of an a.c. generator by Lyapunov's direct method," Int. J. Control, vol. 8, 1968, pp. 131-144.

8. J.L. Willems, "Direct methods for transient stability studies in power system analysis," IEEE Trans. Autom. Control, vol. AC-16, August 1971.

9. ———, "Optimum Lyapunov functions and transient stability regions for multimachine power systems," Proc. Inst. Elec. Eng., vol. 117, Mar. 1970, pp. 795-801.

10. J.L. Willems, Stability Theory of Dynamical Systems. New York : Wiley, 1970.

11. ——— and J.C. Willems, "The application of Lyapunov methods to the computation of transient stability regions for multimachine power systems," IEEE Trans. Power App. Syst., vol. PAS-89, May/June 1970, pp. 795_801.

Figure 1 : Multivariable nonlinear feedback system

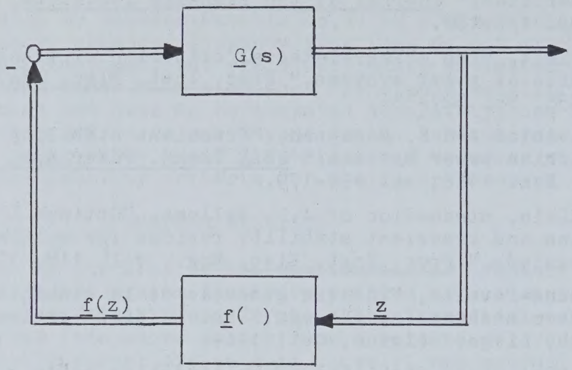
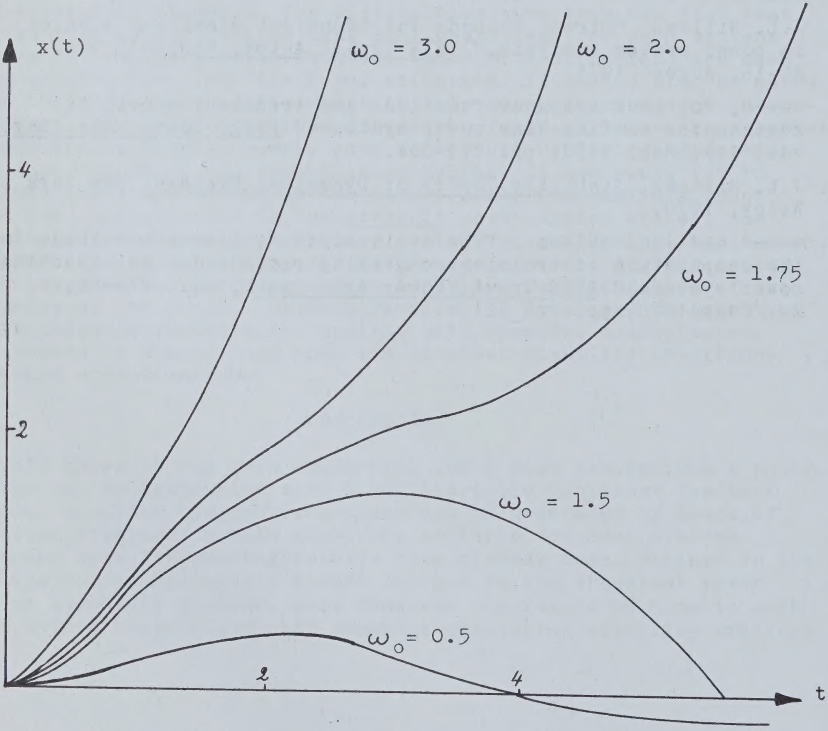


Figure 2 : Trajectories of a non-uniformly damped two-machine system



CRITICAL SURVEY OF TRANSIENT STABILITY STUDIES OF MULTIMACHINE POWER SYSTEMS BY LIAPUNOV'S DIRECT METHOD.

M. RIBBENS-PAVELLA
Department of Electrical Engineering
University of Liège
Liège, Belgium.

ABSTRACT

The variety and the seemingly apparent contradictions which can be found in already published papers dealing with the subject may confuse even specialists. The sole aim of the present survey is to attempt to explain some of the reasons for the existing contradictions in the area.

1. INTRODUCTION

The improvement and spread of non-linear mechanics as well as the increase in the scope of computers and numerical analysis have enabled attempts to rationalise the study of transient stability of power systems. Because of this the use of Liapunov's direct method has given rise to numerous works in the last decade. Two distinct directions of research have opened up in this field : the first is the increasingly close study of the effects of regulators (of speed and voltage) in simple cases (systems of one machine connected to an infinite bus and 2-machine systems) [1]-[7]; the other is concerned with establishing a general theory (multimachine systems) [8]-[22]. Although the first is undoubtedly useful, the second is the one which offers by far the greatest practical interest. This paper only includes discussion of works dealing with multimachine systems.

The present critical study is justified by the large number and diversity - or seeming diversity - of works on the subject. This variety is partly due to the fact that the subject is treated by scientists working in fairly diverse fields and whose aims are not always the same: some incline towards the physical side while others are interested in the mathematical side. The complexity of the physical problem, the above mentioned variety and the use of increasingly sophisticated mathematical tools constitute possible sources of errors and make it difficult to compare certain papers and to estimate their contribution.

Before getting down to the subject we feel it would be useful to state the physical problem as clearly as possible, to point out the standard admitted simplifications in transient stability studies and to indicate the principle of the application of Liapunov's direct method in this domain. These premises can most certainly be regarded as well known and we should like to apologize for this to the specialists; however, experience has shown that quite a few misunderstandings arise from the failure to grasp the elementary considerations which are developed in paragraph 2 and 3.

2. THE PHYSICAL SYSTEM AND LIAPUNOV'S DIRECT METHOD

2.1. We are concerned with a n-machine power system. In its steady state operation, the mechanical input power to the generators is equal, when neglecting losses, to the electrical power delivered by these generators to the network. When a large disturbance or a fault (short circuit...) occurs, this equilibrium is destroyed; the system's motion is then described by n equations of form (see nomenclature) :

$$M_i \ddot{\delta}_i = P_{mi} - P_{ei} \qquad (1)$$

where P_{ei} includes coupling terms between machines (see paragraph 3).

The transient stability problem is the study of the system's capacity to recover, after a disturbance, a synchronous steady state operation (†). Its study is extremely complicated, for the rigorous solution of (1) would necessitate taking into account a large number of parameters which vary in time and are interdependent. Among other things, this study would have to consider the action of voltage and speed regulators which would lead to use Park's equations and to the consideration of variation in mechanical powers; it would also have to consider damping whose rigorous introduction not only complicates the system (1) but also includes parameters difficult to evaluate practically [23].

A first approach to the problem consists of assuming that the mechanical input powers remain constant during the whole disturbance : this is the standard procedure of transient stability studies, namely with the step-by-step method.

2.2. Whatever the approximations used may be, the transient stability problem may be formulated in the following manner : given a power system initially in a steady operation and assuming a disturbance appearing at time t_o :

- is there a stable equilibrium position for the system after the disturbance is cleared and if yes,
- which is the critical clearing time, that is the maximum time that the perturbation may remain without the system loosing its capability to come back to a steady state.

So, whatever the method used may be, the transient stability problem has two steps :

step A which studies the evolution of the system from t_o to the clearing time t_e;

step B which studies the evolution of the system in its post-fault (final) phase :

- if the system is found to be unstable, t_e is superior to the critical clearing time t_c,
- if the system is found to be stable, $t_e < t_c$; obviously, it is the critical clearing time t_c which must be determined by several trials of assumed values of clearing times t_e.

Both of the two above steps are studied in the standard numerical method by numerical integration of the system's equations of motion, and computation of the swing curves (i.e. the curves of the evolution of the rotor angles) of the system in its final phase (step B). Due to the assumption of constant mechanical input powers during transients, the system will be considered as stable if the swing curves of all the generators, although not parallel, do not diverge indefinitely with respect to each other. (Note that the mean frequency of the system after disturbance is generally slightly different from its nominal frequency). Obviously, the computation of t_c by the standard step-by-step method is quite long.

---

† More precisely, the transient stability is defined as follows : "An electrical power system is said to be in a transient stability state with respect to a large perturbation if, after such a perturbation, it recovers a synchronous operating state" (Conference Internationale des Grands Réseaux Electriques à Haute Tension, "CIGRE", Paris, 8-18 June, 1966)

2.3. The principle of Liapunov's direct method when applied to the transient stability studies consists of the replacement of the above mentionned step B by a stability criterion; this criterion is based on the construction of a Liapunov function V in order to determine in the state space the stability domain surrounding the stable solution of the post-fault system (system in its final stage (†). This domain is bounded by $V_\ell = V(\delta^u, 0)$ where V_ℓ is the smallest of the values that V takes in the unstable equilibrium points surrounding the stable one and where $(\delta^u, 0)$ is precisely this point. Conditions (δ^e, ω^e) are computed for several clearing times t^e by numerical integration of the faulted system equations (step A) and $V(\delta^e, \omega^e)$ are compared to V_ℓ, until $V(\delta^c, \omega^c) = V_\ell$: t^c is the critical clearing time. It is obvious that the success of this method will depend on :

- the quality of the Liapunov function : it is important to construct the most suitable Liapunov function for the physical problem under consideration, which must possess above all the largest possibly stability domain.
- the speed of numerical computation of the system's equilibrium points in its final stage : these are solutions of the non-linear equations obtained from the motion equations when the mutual n machines' accelerations are reduced to zero.

Keeping in mind the above elementary considerations we will now establish the mathematical model.

3. MATHEMATICAL MODEL

3.1. Let us consider an n-machine electrical power system without damping effects. The motion of the i^{th} machine, when a fault occurs, is described by an expression derived from (1) :

$$M_i \ddot{\delta}_i = P_{mi} - E_i^2 G_{ii} - \sum_{\substack{j=1 \\ j \neq i}}^{n} A_{ij} \cos(\delta_i - \delta_j - \theta_{ij}) \quad (2)$$

In this expression it is assumed that the only variables are the rotor angles δ_i, δ_j. In fact, this assumption as well as the formulating of (2) involve the adoption of the simplifying hypotheses, currently admitted when studying transient stability, i.e. :

- instantaneous electrical power expressed during the disturbance by a relation of the same form as the active power in steady state;
- generator represented by its internal voltage behind its transient reactance of constant modulus E, in series with its transient impedance which is also constant;
- constant inertia coefficient of each generator ($M_i = \omega_o I_i$; I_i : inertia moment); this assumption is justified because the speed variations occuring in practice are very small (a few per cent);
- constant terms of the nodal admittance matrix of the network during the disturbance; in addition this matrix is symmetrical (††);

---

† It is to be noted that such a solution may be non-existant in the case of very large disturbance.

†† Except in the case of networks including transformers with complex ratios [24].

- constant mechanical input power to each machine during the disturbance (see paragraph 2).

Moreover, the construction of a Liapunov function in the multimachine case has only been possible until now with the additional assumption :

$$\theta_{ij}=\frac{\pi}{2} \quad (i,j=1,2,\ldots n;\ i\neq j)\ . \tag{3}$$

This leads to the neglect of the conductances of the transfer admittances, that is *approximately* the resistances of the connections among the generators; this assumption comes very close to the physical reality in most cases. In the general case such an assumption must logically lead to a somewhat conservative stability criterion. Anyway, as far as we know, nobody has been able to construct a correct Liapunov function without this simplification.

With the above assumptions the absolute motion of each machine will be described by a differential equation of the form:

$$M_i\ddot{\delta}_i=P_{mi}-E_i^2G_{ii}-\sum_{\substack{j=1\\ j\neq i}}^{n}A_{ij}\sin(\delta_i-\delta_j)=P_i-\sum_{\substack{j=1\\ j\neq i}}^{n}A_{ij}\sin(\delta_i-\delta_j)$$
$$(i=1,2,\ldots.,n). \tag{4}$$

However, it comes from the foregoing that the study of the *whole* has to be made not by means of the *absolute* rotor angles, δ_i,δ_j, but of the *relative* ones $\delta_{in}=(\delta_i-\delta_n),\delta_{jn}=(\delta_j-\delta_n)$, where n acts as a comparison variable (†). Indeed, the important thing here is that the system must be able to be kept in step before the governors which will bring it back to a steady state at nominal frequency are entered into action. Now, the keeping in step will be achieved if the *difference* between the rotor angles with respect to any of them taken arbitrarily as a comparison angle remains constant or almost constant, even if the speeds are different from the synchronous speed and if the accelerations are different from zero [21]. This means that the mathematical model must be described not by n equations of form (4) but by (n-1) equations deriving from the preceding ones and with the variables δ_{in} (i=1,2,....,n-1).

On the basis of the above considerations the general expression of the motion of the disturbed system can be written as follows :

$$\ddot{\delta}_{in}=D_{in}-A_{in}(M_i^{-1}+M_n^{-1})\sin\delta_{in}$$
$$+\sum_{\substack{j=1\\ j\neq i}}^{n-1}[M_i^{-1}A_{ji}\sin(\delta_{jn}-\delta_{in})-M_n^{-1}A_{jn}\sin\delta_{jn}]\equiv f_{in} \tag{5}$$
$$(i=1,2,\ldots,n-1)$$

or :

$$\dot{\delta}_{in}=\omega_{in} \qquad \dot{\omega}_{in}=f_{in}\ . \tag{6}$$

Hence, the system's motion is described in a 2(n-1)-dimensional state space, with the state variables $(\delta_{in},\omega_{in})$; its equilibrium points are given by :

$$\omega_{in}=0,\quad f_{in}=0\ . \tag{7}$$

---

† It would be equivalent to consider the generators'motion with respect to the mean motion of the whole system defined by [22] :

$$\overline{\delta}=\frac{1}{\Sigma M_i}\delta_i\ ;\ \overline{\omega}=\frac{1}{\Sigma M_i}\Sigma M_i\omega_i \qquad (i=1,2,\ldots.,n)$$

REMARK.

The equation (5) shows that it is not necessary to consider as constants the mechanical input powers P_{mi}; these powers *can vary as long as their variation is proportional to their respective inertia coefficients*. Indeed, from D_{in} = constant, we get :

$$\Delta D_{in} = 0, \text{i.e. } \frac{\Delta Pmi}{M_i} = \frac{\Delta Pmn}{M_n} = C.$$

This restriction is weaker than the constance of the P_{mi}[11].

Particular case:System with an "infinite" bus.

Let n be this "infinite machine" : $M_n = \infty$. The general expression (5) is then transformed into :

$$\ddot{\delta}_{in} = M_i^{-1}[P_i - A_{in}\sin\delta_{in} + \sum_{\substack{j=1 \\ j\neq i}}^{n-1} A_{ij}\sin(\delta_{jn}-\delta_{in})] \equiv f_{in} . \qquad (8)$$

Obviously, equations (6) and (7) remain valid on condition that the suitable expression of f_{in} is used.

3.2. System with damping torques.

The influence of damping may be *approximately* described by means of a term in $\dot{\delta}_{in}$; in this case equation (5) is to be replaced by :

$$\ddot{\delta}_{in} = f_{in} - C\dot{\delta}_{in}$$

where

$$C = \frac{C_i}{M_i} = \text{constant } (\dagger) . \qquad (5')$$

Expression (6) is replaced by :

$$\dot{\delta}_{in} = \omega_{in}$$
$$\dot{\omega}_{in} = f_{in} - C\omega_{in} \qquad (6')$$

whereas expression (7) remains unchangeable.

3.3. Comments.

Let us insist on the fact that, given the admitted assumptions in paragraph 3.1, the consideration of system (4) as a mathematical model in transient stability studies is wrong both from the mathematical and from the physical point of view. If necessary, the following remarks indicate this statement clearly.

From a mathematical point of view, system (4) is visibly unstable. Indeed, if δ_i^o (i=1,2,...n) is a solution of the ordinary differential equation (4) then $(\delta_i^o+\alpha+\beta t)$ will also be a solution (α,β : any constants but the same for all the machines (††) : in this case the rotor angles would be able to increase indefinitely with time. This is obviously absurd.

From a physical point of view, the statement according to which the system's equilibrium points are given by reducing to zero the right hand members of (4), that is by :

$$P_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} A_{ij}\sin(\delta_i - \delta_j) = 0 \quad (i=1,2,....,n) \qquad (9)$$

---

† In the case of uniform damping torques as it is considered here the number of state variables remains 2(n-1); if not, this number is (2n-1).

†† It should be noticed that β cannot be very large because if it is, then the approximation of constant terms of the admittance matrix no longer applies; indeed β represents here the speed"above" the synchronous speed of the system which Y's terms depend on.

is also wrong. Indeed by adding the n expressions (9) we obtain:

$$\Sigma P_i = 0. \qquad (10)$$

But this relation expresses the system's energy conservation without any variation of kinetic energy. This condition is obviously required for the normal load-flow solution of a system operating with a rigorously constant frequency. It is not required for the equilibrium of a disturbed system (it is even highly improbable). In other words, this would imply that one should consider as unstable a network whose frequency is not exactly the nominal frequency (†).

The above considerations are perfectly well known by engineers in electrical power systems stability (e.g. see Dahl, 1938 [25] and Kimbark, 1948 [26]).

In addition, Aylett who, due to his excellent paper published in 1958 [27] may be regarded as the forerunner of the application of Liapunov's direct method to the transient stability study, used the correct mathematical model. It is much more curious to see relations (4) used in transient stability studies as the mathematical model and relation (10) taken into account implicitly or explicitly in several papers dealing with the subject [8-10,13,17,18,20]. It is also strange to find in the same paper the treatment of a 2-machine system made from one ordinary differential equation whereas the 3-machine system is treated from 3 ordinary differential equations.

On the contrary, some other authors, namely Di Caprio and Saccomano [11] [19], Tagirov [14], Putilova [15][16], Pal [28], Morshkin [29] and the author of the present paper [12][21][22] have pointed out this frequent mistake.

To be sure, the construction of Liapunov functions becomes easier when the system of n equations (4) is used. However, this can in no case be regarded as a valid reason.

The above mentioned mistakes disappear and everything goes back to normal when the system has a machine connected to an "infinite" bus. Indeed, in this case, assuming the n^{th} machine to be the "infinite" bus, we have $\dot{\delta}_n = 0$, δ_n = constant : therefore the angles present in (4) can be regarded, except for a constant, as the relative ones and the number of equations (4) becomes (n-1) (i=1,2,....,n-1). However, this is only one particular case.

3.4. Summary.

The implications of the above considerations are :

- the motion equations of the mathematical model must take into account the relative rotor angles and not the absolute ones; the number of state variables are 2(n-1) (at least when non uniform damping torques are not considered).
- the most suitable Liapunov function (at least when damping torques are not considered) is a function whose derivative is zero everywhere, so as to ensure the simple system's stability; this leads quite naturally to the construction of a Liapunov function by means of a first integral of the system.

---

† In the present case the correct formulation of the system's energy conservation before the governors enter into action is expressed by :

$$\Sigma P_i = \frac{d}{dt} \varepsilon$$

where ε is the kinetic energy stored in the machines.

In addition, it is to be noted that the most rigorous way in taking into account damping due to asynchronous torques, (which are the most important ones), when acting to the i^{th} machine is the term [23],[25] : $\sum_{i=1}^{n} D_{ij}(\dot{\delta}_i-\dot{\delta}_j)$ where D_{ij} are assumed to be constant ($D_{ij} \neq D_{ji}$). Unfortunately, it is not possible for the present time to take into account such damping torques when constructing Liapunov functions.

4. LIAPUNOV FUNCTIONS

4.1. The examination of the papers dealing with the subject shows that the proposed Liapunov functions, when correctly constructed are similar to each other [11][12][15][22].

4.1.1. Liapunov function when damping torques are neglected

As a basis for the following discussion we give, without details, the function we have already proposed(†). This function is constructed from the mathematical model (6) in the 2(n-1)-dimensional state space $(\delta_{in},\omega_{in})$; the origin of this space is situated at the point $(\delta^s,0)$ which is assumed to be a stable solution of (6), given from (7). We get :

$$V(\delta,\omega)=\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}\frac{1}{2}M_iM_k(\omega_{in}-\omega_{kn})^2$$

$$-\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}[B_{ik}(\delta_{in}-\delta_{kn})+A_{ik}\cos(\delta_{in}-\delta_{kn}).M]+K$$

$$=A(\omega)-\Phi(\delta)+K \qquad (11)$$

with : $K=\Phi(\delta^s)$.

We have proved that the function V, in a certain region Ω of the 2(n-1)-dimensional state space, meets the conditions required by Liapunov functions to assure the simple stability of the system (it is called simple to distinguish it from the asymptotic one), that is :

(a) $V(\delta^s,0)=0$

by a suitable choice of K

(b) $\dot{V}(\delta,\omega)=0$

by construction

(c) $V(\delta,\omega)>0$

in a certain domain Ω surrounding the origin, insofar as the hessian of $\Phi(\delta)$ is negative at the origin.

The stability region is found to be bounded by the closed surface

$$V_\ell=V(\delta^u,0)$$

where V_ℓ is the smallest of the values that V takes in the unstable equilibrium points surrounding the stable one; $(\delta^u,0)$ is precisely this point. It should be pointed out that the above Liapunov function is the best one to ensure the simple stability of the system, as its derivative is zero everywhere.

Particular case : System with an "infinite" bus ($M_n=\infty$). V is simplified and becomes :

$$V(\delta,\omega)=\sum_{i=1}^{n-1}(\frac{1}{2}M_i\omega_{in}^2-P_i\delta_{in})-\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}A_{ik}\cos(\delta_{in}-\delta_{kn})+K. \qquad (12)$$

---

† For the details of the construction of V as well as its theoretical treatment (proof that it really is a Liapunov function) see [12] and [21].

4.1.2. Liapunov function when damping torques are included. When based on the mathematical model (5') it is possible to take as a Liapunov function the V function of form (11).

The stability region is still bounded by the same surface V_ℓ used in the case without damping.

However in this case the total derivative of V is no longer zero but becomes a negative semidefinite function; the system's stability is shown to be reinforced in this case (asymptotical instead of simple).

4.2.1. Di Caprio and Saccomano [11],[19] base their study on correct physical statements and consider the 2(n-1) correct state variables δ_{in}, ω_{in}.

The authors replace our condition (3) by another seemingly less restrictive (see in [11], relations (3) and (5)) :

$$\frac{\partial P^+_{ei}}{\partial \delta_{jn}} = \frac{\partial P^+_{ej}}{\partial \delta_{in}} \quad \text{"for every pair of indexes } i,j=1,....,n-1\text{".}$$

But by this one is led to equate to $\pi/2$ all the θ_{ij} successively, which, in spite of appearances comes back to the condition (3). Hence, the mathematical model used becomes equivalent to (5), and the Liapunov function proposed identical with (11) (except for a constant). On the whole, papers [11] and [19] are a good theoretical work, including a thorough study of the constructed Liapunov function.

4.2.2. The papers of Tagirov [14] and Putilova [15]-[16] are exhaustive theoretical as well as experimental studies. The physical statements are properly developed and the mathematical model correctly established.

Instead of replacing in expression (2) all the θ_{ij} $(i \neq j)$ by $\pi/2$, Putilova [15] starts from the correct equation :

$$M_i \ddot{\delta}_i = P_i - \sum_{\substack{i=1 \\ i \neq j}}^{n} A_{ij} \cos(\delta_i - \delta_j - \theta_{ij}) =$$

$$= P_i - \sum_{\substack{i=1 \\ i \neq j}}^{n} A_{ij} [\cos(\delta_i - \delta_j)\cos\theta_{ij} + \sin(\delta_i - \delta_j)\sin\theta_{ij}]$$

and adopts the approximation :

$$M_i \ddot{\delta}_i = P_i - \sum_{\substack{i=1 \\ i \neq j}}^{n} A_{ij} \sin(\delta_i - \delta_j)\sin\theta_{ij} \ .$$

From the above expression the author constructs the mathematical model and a Liapunov function which, by this means, takes partially into account the transfer conductances. For $\sin\theta_{ij} = 1$ this Liapunov function coincides with (11) (except for a constant).

In addition to the stability criterion provided by comparing the values $V(\delta^e, \omega^e)$ and $V_\ell(\delta^u, 0)$, Tagirov [14] proposes a non-stability criterion; this is due to the fact that the Liapunov stability criterion provides somewhat restrictive results (†). The latter should be defined, according to Tagirov, by :

(†) There are three reasons for this [22]:

Firstly, it is known that in practice the above stability criterion provides sufficient but not necessary conditions (for systems of more than two machines). Indeed, when the system's motion is represented in the state space by a point situated in-

$$|x_i(t)| \geq \max_i |x_{ic}|$$

where x_{ic} are the coordinates of the point δ^u and where $x_i(t)$ are the coordinates of the disturbed system's motion.

As with all pragmatic criterion, only experimenting can provide information about the quality and the usefulness of the above criterion.

4.2.3. Transient stability studies are generally based on the assumption that the machines' inertia coefficients are constant. In paper [22], these coefficients are no longer regarded as constant but as varying with angular speeds : $M_i=(\omega_o+\omega_i)I_i$ (where I_i is the inertia constant).

In this case, the mathematical model as well as the number of state variables are modified (2n-1 state variables : δ_{in}, ω_i; $i=1,2,\ldots n$).

The authors construct a function and prove that it is a Liapunov function. The total derivative of this function is identically zero when neglecting damping torques and becomes negative semidefinite when including damping torques.

It is interesting to note that in this paper the damping torques acting to the i^th^ machine are described by the terms :

$$\sum_{j=1}^{n} D_{ij}(\dot{\delta}_i-\dot{\delta}_j)$$

Unfortunately calculations are possible only when admitting the condition $D_{ij}=D_{ji}$. In spite of this restriction the consideration of damping torques by means of the above expression is more correct than that used in (5') which, as matter of fact, is a particular case of the latter.

4.2.4. The first trial of the application of Liapunov's direct method to the transient stability of multimachine power systems was that of El-Abiad and Nagappan published in 1966 [9]. The merit of those authors is above all to have been the pioneers in this direction; nevertheless the following remarks and objections must be formulated.

1° *In the general case* the study is made false by the use of the equations (4) as the motion equations (study of the evolution of the absolute rotor angles and hence consideration of 2n state variables δ_i, ω_i; $i=1,2,\ldots n$) : this leads to a wrong mathematical model and to the construction of a function which is not a Liapunov function.

2° *In the particular case* where the system has an infinite bus, i.e. a machine whose inertia coefficient is "infinite", $M_n=\infty$, the above 2n state variables are reduced to 2(n-1) : the

---

side the limit surface V_ℓ, the system is certainly stable; but it may still remain stable even if it is represented by a point located outside V_ℓ.

Secondly, it is also known that the above stability criterion does not include the time explicitly. This means that this criterion does not provide any information about the time after which a system, when unstable, will actually start being out of step; but, from a practical point of view, a system which does keep in step during the first two or three oscillations may be regarded as stable.

Thirdly, the stability criterion may be even more conservative because of the neglect of the transfer conductances in the mathematical model.

absolute rotor angles are, implicitly, replaced by the relative ones and the state variables automatically become δ_{in}, ω_{in} $(i=1,2,\dots,n-1)$. In this case two alternatives must be considered :

- Transfer conductances are neglected (see our condition (3)) : the motion equations are of form (8) and everything goes back to normal. The mathematical model becomes correct and the function constructed by means of it is a real Liapunov function and coïncides with (12).
- Transfer conductances are taken into account : the function constructed is no longer a Liapunov function, for its total derivative is not a negative definite or semidefinite function. In this case the authors propose a stability criterion in "the sense of Lagrange". However such a criterion has no practical interest; indeed, according to the authors' criterion a system should be considered stable even if some of the relative rotor angles increase indefinitely.

To summarize, the function proposed in [9] is a real Liapunov function only in the particular case of a power system including an infinite bus and provided that the transfer conductances are neglected.

The numerical example treated by the authors with a 4-machine system gave results agreeing with those of the standard procedure, probably because this system has a machine much heavier than the others ($M_1 \simeq 36M_3 \simeq 50M_1 \simeq 50M_4$, see in [9], p.173, Table II); indeed in this case we are concerned practically with a system including an infinite bus.

4.2.5. The papers [13] and [17] use a theorem of Anderson to prove the generalized Popov criterion for feedback with multiple non-linearities. Paper [17] may be regarded as the extension and the systematic application of ideas presented in [13]; that is why we will examine them together.

The mathematical tools used are very sophisticated and the method seems very interesting. But the motion equations and the state variables considered are wrong. Indeed the authors take into account equations of form (4) and note that these n equations are related by the condition (10); they conclude that the number of the state variables is 2n-1. This is wrong as we indicated in paragraph 3.

By using Anderson's theorem, the author constructs in [17] a family of functions obtained from a given expression when a parameter μ is varied. For two particular values of μ ($\mu=0$ and $\mu=-1/M$) he obtains two particular functions, $V_1(x)$ and $V_2(x)$ respectively. $V_1(x)$ is the El-Abiad and Nagappan's function. $V_2(x)$ is *formally* equal to V/M where V represents the function (11). It is to be noted that $V_2(x)$ becomes a real Liapunov function when considering the 2(n-1) correct state variables δ_{in}, ω_{in} rather than the 2n-1 ones taken into account by the author.

In the particular case where $M_n = \infty$, the functions $V_1(x)$ and $V_2(x)$ coïncide (except for the constant M) and become real Liapunov functions, as the author himself points out. This may be explained by the fact that in this case one is implicitly led to consider the 2(n-1) correct state variables.

Furthermore, paper [17] includes a comparative stability study of systems with and without damping torques; the same remarks as above must be applied to this study.

4.2.6. Paper [18] uses a technique similar to that of the papers [13] and [17] though the mathematical developments in constructing functions V are different. The generation of these functions is based on Popov's method and works of Kalman and Narendra and Neuman.

Remarks similar to those of paragraph 4.2.5. can be made about paper[18].

The motion equations used here are of form (4) and the number of state variables considered is 2n : δ_i, ω_i, i=1,2,...,n (see in [18], p.3, equation (19) for the 3-machine system).

The authors construct families of functions. For particular values of certain parameters, one finds again a function V(x) *formally* similar to ours, (11), but which is not a real Liapunov function in the 2n-dimensional state space. (This becomes evident when, for instance, replacing in equation (37), p.5 of[18], ω_1 by ω_1+c, ω_2 by ω_2+c, ω_3+c : the value of V(x) remains unchangeable; obviously, this is no longer valid when one considers the variables ω_{13} and ω_{23}). Obviously this function V(x) automatically becomes a real Liapunov function when using the correct state variables or when $M=\infty$.

4.2.7. The paper [20] includes a large number of considerations but also some inaccuracies.

In section II of his paper, the author starts by studying two functions : the V_1 (which is the El-Abiad and Nagappan function) and the V_2 (which is a function proposed by the author himself). To do this he establishes a mechanical analog of a power system by comparing their respective motion equations. But the motion equations considered for the power system are of type (4); this is wrong. The same objection applies to the mechanical system which, such as considered by the author, does not possess any stable equilibrium point (the reference axis is by no means tied to the rest of the system). Therefore the technique used in IIC to prove that V_1 is a Liapunov function is not valid. In addition, in order to prove that V_2 is a Liapunov function in a n(n-1)-dimensional state space, the author uses the fact that V_1 is a Liapunov function in a 2n-dimensional state space; this also cannot apply. In fact, it is possible to prove that V_2 is a Liapunov function provided that the state variables are correctly chosen (δ_{in}, ω_{in} in the 2(n-1) - dimensional state space) and that the transfer conductances are neglected. In addition V_2 must be correctly formulated (see on p.24, expression (3), where the right-hand terms must be added and suitably grouped before being integrated).

In section III the author proposes three new functions. Again, we do not find any proof of the fact that they are Liapunov functions.

To summarize, we can say that among all the functions proposed by the author the only Liapunov function is the V_2 provided that the above remarks are taken into account. V_2 coincides in this case with the function (11).

Let us notice that the paper includes some contradictions. (For instance, see the expressions (20) and (30) where one can read respectively $M_i\ddot{\delta}_i=0$ (which is false) and $M_i\ddot{\delta}_i = \text{constant} \neq 0$ (which is true).

In spite of these theoretical imperfections, the author of this paper has the merit of having treated numerically a 10-machine system.

4.2.7. The papers [8] and [10] use and compare each others' different standard techniques of generation of functions V. However, the mathematical model as well as the number of the state variables are not correct (2n state variables except in [7], equation (40) where one considers n(n-1) state variables).Under such conditions, comparisons and discussion become hardly possible.

4.2.8. The technique of generation of Liapunov functions presented in [6] is completely different. It is based on Zubov's method and applied to the particular case of a synchronous machine connected to an infinite bus. The Liapunov function is constructed in a series' expansion with a theoretically infinite number of terms; this seems quite normal as the expected function will include trigonometrical expressions.

The author indicates that in the particular case of a machine connected to an infinite bus, a highly truncated Liapunov function series is needed to approximate the stability boundary. He notices that this method can be generalized in the multimachine case but does not give any directions about this; he only indicates that in this general case the generation of Liapunov functions requires large computational time. So at the present time, we don't have the elements necessary to appreciate and to compare the behaviour of functions generated by this means with function (11).

4.3. Conclusions. With the assumptions admitted in paragraph 3.1. and when the damping torques are neglected, the correct mathematical model for the system's stability study is provided by (5) and (6) or any other equivalent variant. Under these conditions the Liapunov function (or more precisely the most suitable Liapunov function for the transient stability study) will be given by (11) or any other expression equivalent to (11). This will hold good, whatever the mathematical tool used for its construction may be and however simple or sophisticated it is.

5. PRACTICAL USE OF THE LIAPUNOV CRITERION.

5.1. Contrary to the large number of theoretical works, only a few papers including numerical examples have been published until now [9],[12],[16],[20] and [22].

The essential problem which must be solved in order to make the Liapunov stability criterion work in practice is the computation of the solutions of the system of non-linear equations (7). These solutions are periodic, with periodicity 2π; thus it is a question of finding only the solutions in an interval 2π. As a matter of fact, among all these solutions only two really intervene in the formulation of the stability criterion :

- the stable solution $(\delta^s,0)$
- the unstable solution $(\delta^u,0)$.

Let us at once note that for the reasons we have already pointed out [12] [21], we do not need to know these solutions with great accuracy. This remark allows a gain on computational time. Let us also note that the computation of the stable solution is in general fast (†), for it is located in the neighbourhood of the stable solution of the pre-fault system, which is given by the load-flow program preceding the stability program.

---

† Except in the case of very great disturbances, as, for instance, the elimination of a high power machine or of a line transporting a large amount of energy; in such cases it could happen that no solution at all exists anymore.

Unfortunately this does not apply to the unstable solution δ^u, whose determination represents the most important computational operation for the application of the Liapunov stability criterion. Indeed, until now there is no criterion which enables us to predetermine, among all the unstable solutions surrounding the stable one, the solution which provides the value V_ℓ. Furthermore, the number of these singular points depends not only on the number of the system's machines but also on the particular behavior of its parameters (†).

The above discussion shows that the essential problem to be solved in order to make the stability criterion work satisfactorily is the determination of V_ℓ, i.e. of δ^u.

5.2. The experiments made in the Institute of Energetical Research of Novosibirsk by Putilova and her colleagues [16] are very interesting for many reasons. These calculations were based on power systems of 24, 15, 12, 5 and 1 machines. Their goals were numerous as explained below.

1° Comparison of the results provided by the 24-machine system with those obtained by the reduced 15-machine, 12-machine, 5-machine and 1-machine systems. The results obtained via the stability criterion coïncide with those obtained via the step-by-step method in the case of 1-machine system and are somewhat conservative in the multimachine case; this conclusion is in accordance with the theoretical considerations (see paragraph 4.2.2.).

2° Comparison of the results provided via the stability and the non-stability criteria [14]. The experiments show that the first ones are somewhat conservative (see 1°), whereas the latter are too optimistic when compared to the results of the step-by-step method.

3° Comparison of the values of the V_ℓ computed via the unstable point, δ^u, solution of the non-linear system (7), and of the V'_ℓ provided by an approximate expression proposed in [16] (method of the "n-dimensional cube"). The latter does not require the computation of the solutions of (7) and could be shown to be very advantageous for a first approximate value of V_ℓ; indeed, the experiments show that the stability criterion provided via the V'_ℓ value is quite close to that provided by the V_ℓ value.

4° Moreover, the stability criterion is found to be in a better accordance with the standard numerical method when taking into account a greater number of oscillations. Again, the results of these two methods coïncide in the case of 1-machine system, for in this case the non stability, when it occurs, actually takes place during the first oscillation. These considerations are quite obvious from a theoretical point of view.

The results of the above studies are very interesting and make paper [16] worthwhile.

5.3. In the paper [12] two power systems were considered : a 3-machine system and a 9-machine system. The stability criterion was checked by means of two Liapunov functions : the function V, given by (11), and constructed from strict mathematical calculations obtained when all the $\theta_{ij}=\pi/2(i,j=1,2,\dots,n;i\neq j)$ and the function V_2 constructed from approximate calculations when only certain θ_{ij} are taken as equal to $\pi/2$. The comparison of the results thus obtained with those obtained by the step-by-step method shows that in the 3-machine system V_2 gives slightly better results than V (†) whereas in the 9-machine case V and V_2 give the same result, which is somewhat conservative (††). It is to be noted that the study of a 3-machine system enables us to give a planar representation of the curves V=constant (plane ω=constant); these curves are found to agree fairly well with theoretical consideration.

---

† In our paper [21] we have noticed that by physical considerations one can find 2^{n-1} starting points to determine the equilibrium points of the post-fault system; however the number of solutions of (7) is not always 2^{n-1} as has been shown in [30].

The computation of the solutions has made by the Davidon-Fletcher-Powell method which, as far as we know, is the most appropriate to the form of the equations (7) and seems to be satisfactorily rapid (†††).

5.4. The first to have used the stability criterion in a realistic case are El-Abiad and Nagappan [9]. These authors considered a 4-machine system and evaluated the critical clearing time for a 3-phase fault. The comparison of the results obtained by the Liapunov criterion with those obtained by the standard step-by-step method are shown to be in very good agreement. As we have already noticed this is probably due to the fact that one of the considered system's machines is much heavier than the others so that in this case we can practically consider that one of the machines is "infinite".

5.5. Paper [20] takes into account a 10-machine power system; it includes ideas and useful practical considerations, namely some concerned with the reduction of the number of machines in order to reduce the number of non-linear equations. The critical clearing time obtained by the Liapunov stability criterion although somewhat conservative seems in a good agreement with that given by the step-by-step method. A comparison of the computational time required by each of the two above methods leads the author to the conclusion that the first one needs only 60 per cent of the computing time required by the second.

5.6. Paper [22] considers a 9-machine system in order to compare the influence of the variation of the M_i on the system stability. The results would indicate a slight tendency to weakening the stability; however this result may possibly be invalidated by the fact that the step-by-step program providing the conditions (δ^e, ω^e) used in the computation of V, does not take into account the variation of the M_i; consequently the authors insist that this result must be regarded with caution.

6. CONCLUSIONS

The application of Liapunov's direct method to the transient stability of multimachine power systems has stirred up a lot of discussion during the last decade. This is because this method is very attractive both from the theoretical and from the practical point of view. From a theoretical point of view the establishment of a rational criterion is obviously satisfactory. From a practical point of view the method offers several advantages over the standard step-by-step method; the most important are the determination of a "stability index" of a given system [31] and an expected gain on computational time which will permit on-line calculations in transient stability studies.

---

† t_{c2} = 0.312 s; t_c = 0.321 s; $t_{c\ actual}$ = 0.3 s.

†† $t_{c2} \simeq t_c$ = 0.43 s; $t_{c\ actual}$ = 0.44 s.

††† For instance in the case of the 9-machine system the stable solution was found in 0.5 s. with an IBM 360/65.

However, sometimes theoreticians can be so attracted by such satisfactory methods that they are blinded to the basic physical phenomenon. In addition the experimental point of view which is much more tedious has often been neglected. However in this field experimenting, even on a small scale, is indispensable for two reasons :

- on one hand in order to verify the admitted simplifying assumptions; indeed the physical problem is so complicated that it is impossible to approach it without simplifying it first; it is the experiment which enables us to confirm or to invalidate the merit of these assumptions;
- on the other hand to make this method work in practice, namely by resolving numerical analysis questions; besides, the experiment has the advantage of finding and therefore avoiding possible theoretical mistakes.

In any case, due to the number and quality of works carried out in this field, the first steps towards the practical application of the method have already been taken by now : at the moment we can no longer look on Liapunov's direct method applied to power systems as a purely speculative element of only academic interest. However, many improvements still have to be carried out to make the method faster and to free it from many assumptions.

NOMENCLATURE

M_i : inertia coefficient of the i^{th} machine.

$M = \sum_{i=1}^{n} M_i$ (n : number of the system's machines).

E_i : modulus of the internal voltage behind the transient reactance of the i^{th} machine whose argument is δ_i.

δ_i : rotor angle between the internal voltage $\overline{E}_i$ and an axis rotating at synchronous speed ω_o.

ω_i : difference between the speed of the i^{th} machine and the synchronous speed : $\omega_i = \dot{\delta}_i$.

δ_{ik} : angular difference between the rotors of the i^{th} and k^{th} machines : $\delta_{ik} = \delta_i - \delta_k$.

P_{mi} : mechanical input power to the i^{th} machine.

P_{ei} : electrical input power to the network from the i^{th} machine.

$G_{ii} = Y_{ii} \cos\theta_{ii}$, where Y_{ij} is the modulus of the ij^{th} term of the admittance matrix, Y, of a network reduced to the machine ports; θ_{ij} is the argument of the ij^{th} term of Y.

$A_{ij} = E_i E_j Y_{ij} = A_{ji}$

$P_i = P_{mi} - E_i^2 G_{ii}$

$B_{ik} = P_i M_k - P_k M_i$

$D_{ik} = B_{ik}/M_i M_k$

REFERENCES

1. N. Dharma Rao, "A new approach to the transient stability problem", TRANS. AIEE, Vol. 60, June 1962.
2. G.E. Gless, "Direct method of Liapunov applied to transient power systems stability", IEEE TRANS. POWER APPARATUS AND SYSTEMS, vol. PAS-85, n°2, February 1966.
3. F. Fallside, M.R. Patel, "On the Application of the Lyapunov Method to Synchronous Machine Stability", INTERNATIONAL JOURNAL CONTROL, vol.4, n°6, 511-513, 1966.

4. M.W. Siddique, "Transient stability of an a.c. generator by Liapunov's direct method", INTERNATIONAL JOURNAL CONTROL, vol.8, n°2, 131-144, 1968.
5. J.L. Willems, "Improved Lyapunov function for transient power system stability", PROC. IEE, vol.115, n°9, September 1968 (London).
6. Yao-Nan Yu, Khien Vongsuriya, "Nonlinear Power System Stability Study by Liapunov Function and Zubov's Method", IEEE TRANS. POWER APPARATUS AND SYSTEMS, vol. PAS-86, n°12, December 1967.
7. M.A. Pai, M.A. Mohan and J.G. Roa, "Power system transient stability regions using Popov's method", IEEE TRANS. POWER APPARATUS AND SYSTEMS, May/June 1970.
8. N. Dharma Rao, "Generation of Liapunov functions for the transient stability problem", TRANS. OF THE ENGINEERING INSTITUTE OF CANADA, Vol. 11, N° C-3, October 1968.
9. A.H. El-Abiad, K. Nagappan, "Transient stability regions of multimachine power systems", IEEE TRANS. POWER APPARATUS AND SYSTEMS, vol. PAS-85, n°2, February 1966.
10. N. Dharma Rao, "Routh-Hurwitz conditions and Lyapunov methods for the transient stability problem", PROC IEE, Vol. 116, n°4, April 1969.
11. U.Di Caprio, F.Saccomano, "Application of the Liapunov's direct method to the analysis of multimachine power system stability", Third Power Systems Computation Conference, PSCC, Rome, 23-27 Juin 1969.
12. M.Ribbens-Pavella, "Théorie Générale de la stabilité transitoire de n machines synchrones", Doctoral Thesis, University of Liège, 1969. Publications de la revue "Collection des Publications de la Faculté des Sciences Appliquées de l'ULg, fascicule 17.
13. J.L. Willems, J.C. Willems, "The application of Lyapunov direct methods to computation of transient stability regions for multimachine power systems", IEEE TRANS. ON POWER APPARATUS AND SYSTEMS, May/June 1970.
14. M.A. Tagirov, "On the formulation of criteria of transient stability for a mathematical model of power systems", (in Russian). Second Seminary-Symposium on the application of the Liapunov functions method on Energetics; Novosibirsk, 1970. See also : A.T. Putilova and M.A. Tagirov, "Non-stability criteria for multimachine power systems", (in Russian). Reports of the Siberian Institute of Scientific Power Researches; fasc. 17 : "Stability of Power Systems"; Moscou 1970.
15. A.T. Putilova, "Liapunov function for reciprocal equations of motion of synchronous machines and its modification" (in Russian). Second Seminary-Symposium of the application of the Liapunov functions methods on Energetics; Novosibirsk, 1970. See also : A.T. Putilova and M.A. Tagirov, "Liapunov functions for reciproqual equations of motion of synchronous machines", (in Russian). Reports of the Siberian Institute of Scientific Power Researches; fasc.17 : "Stability of Power Systems"; Moscou 1970.
16. A.T. Putilova, "Analysis of complex power systems stability according to Liapunov criterion", (in Russian). Second Seminary-Symposium on the application of the Liapunov functions method on Energetics; Novosibirsk, 1970.
See also : A.T. Putilova, "General Characteristics of Computational Algorithm of Multimachine Systems by Liapunov's Criterion", (in Russian). Reports of the Siberian Institute of

Scientific Power Researches; fasc. 17 : "Stability of Power Systems"; Moscou 1970.
17.J.L. Willems, "Optimum Lyapunov functions and stability regions for multimachine power systems", PROC IEE (London) Vol. 117, n°3, March 1970.
18.M.A. Pai and M.A. Mohan, "Transient stability of multi-machine power systems by Popov's method", IEEE Summer Power Meeting, Los Angeles, July 1970.
19.U.Di Caprio, F.Saccomano, "Non-Linear Stability Analysis of Multi-machine Electric Power Systems", RICERCHE DI AUTOMATICA, Vol.1, N°1, September 1970.
20.G.A. Lüders, "Transient Stability of Multimachine Power Systems via the Direct Method of Lyapunov", IEEE TRANS. ON POWER APPARATUS AND SYSTEMS, Vol. PAS-90, n°1, January/February 1971.
21.M.Ribbens-Pavella, "Transient Stability of Multimachine Power Systems by Liapunov's Direct Method", IEEE Winter Power Meeting, New York, January 1971.
22.E. Carton and M.Ribbens-Pavella, "Liapunov methods applied to multimachine transient stability with variable inertia coefficients" (to be published in the PROCEEDINGS of the IEE).
23.A. Calvaer, Notes du cours de "Stabilité et Réglage des Réseaux Electriques", Université de Liège, (Unpublished).
24.M.Pavella, "Caractères analytiques d'immittances des réseaux polyphasés contenant des transformateurs déphaseurs",BULLETIN SCIENTIFIQUE DE L'A.I.M., Mars-Avril 1967.
25.G.C. Dahl, "Electric Power Circuits, Theory and Applications", Mc Graw Hill, Vol. II, 1938.
26.E.W. Kimbark, "Power System Stability", John Wiley, 1948.
27.P.D. Aylett, "The energy-integral criterion of transient stability limits of power systems", IEEE TRANSACTIONS, July 1958
28.M.K. Pal, "Optimum Lyapunov Functions and Stability Regions for Multimachine Power Systems", Correspondance, PROC. IEE, Vol. 117, n°9, September 1970.
29.Yu.V. Morshkin, "Stability study of power systems".(in Russian). Second Seminary-Symposium on the application of the Liapunov functions method on Energetics; Novosibirsk, 1970.
30.R.J. Thomas, R.D. Barnard, J. Meisel, "The Generation of Quasi Steady-State Load flow Trajectories and Multiple Singular Point Solutions. IEEE Winter Power Meeting, January-February 1971.
31.A. Calvaer, J.L. Horward, M. Ribbens-Pavella, "Applications d'un critère général de stabilité transitoire des réseaux multimachines" (to be published).

POWER SYSTEM STATE REGULATOR DESIGN WITH DOMINANT EIGENVALUE SHIFT

YAO-NAN YU and HAMDY MOUSSA
Department of Electrical Engineering,
University of British Columbia,
Vancouver 8, B.C., Canada.

ABSTRACT

The paper presents a new state regulator design technique of determining the state weighting matrix Q with a dominant eigenvalue shift within controllers practical limit. After a brief account of some changes in concepts of power system stability and difficulties encountered in the earlier state regulator design[1-4], the new design technique is presented and applied to a one machine[5] as well as a multi-machine power systems[6]. It shows that the state regulator provides a better stabilization scheme than others for power systems.

POWER SYSTEM STABILITY PROBLEM

The power system engineering is dealing with a multi-billion dollar business, managed by one of the most experienced engineer groups, yet confronting numerous new challenges due to the fast growing of the system. The system grows not only in size due to interconnection but also in complexity such as HVDC transmission. It also faces new difficulties, e.g., new generator design with less inertia and larger reactances, which implies that the power system stability will rely more heavily upon the control technique in the future.

The power system stability concept has gone through some changes. In the early days, generators are designed with large inertias and smaller reactances, exciter-regulator system provided with stabilizing circuits and governors with dashpots, mainly to improve the performance of the subsystems.

Since most power systems will lose its stability in a matter of a few seconds in case of severe disturbances, and the response of non-electric controllers is rather sluggish, efforts have been made in recent years on the stabilization of power system through excitation control[7-11]. The main objective is to increase the damping and synchronizing torque components of the synchronous machine using phase-compensated speed or power signals.

While the excitation control proves effective for individual power systems with given loads and at certain system oscillation frequencies, there is a definite need of developing new stabilization techniques for large interconnected power systems. While a large number of papers are published on the application of Liapunov's direct method to power system stability analysis and mainly on very low order machines, optimal control papers begin to appear in power literature[12-18]. For the stabilization of large power systems, the most promising one, at the moment, is the state regulator design.

DEVELOPMENT OF POWER SYSTEM STATE REGULATOR DESIGN

It is known also to power engineers that in the state regulator design the optimal control

$$u = -R^{-1}B'K\,y \qquad (1)$$

is found from the minimization of a chosen cost index

$$J = \frac{1}{2}\int_0^\infty (y'Qy + u'Ru)dt \tag{2}$$

constrained to the system dynamics

$$\dot{y} = Ay + Bu \tag{3}$$

One of the difficulties of the state regulator design was the solution of the Riccati matrix K. In our earlier design for power systems Vongsuriya[2] solved K from the matrix Riccati differential equation

$$\dot{K} = KA + A'K - KSK + Q \tag{4}$$

where

$$S = BR^{-1}B' \tag{4a}$$

and numerical instability of integrating n(n+1)/2 equations for an n-th order system was the problem. To apply Puri and Gruver's method[19] of solving

$$KA + A'K - KSK = -Q \tag{5}$$

iteratively in the form of

$$K^{(j)}A^{(j)} + A^{(j)'} + Q^{(j)} = 0 \tag{6}$$

Wedman solved $K^{(j)}$ from an integral solution of the Liapunov type equation (6) using a truncated series[3] similar to Davison and Man[20]. An initial $K^{(o)}$ must be found so that the closed loop system is initially stable[19]. Wedman found the $K^{(o)}$ from an eigenvalue shift using the sensitivity coefficient of eigenvalues with respect to the elements of K, λ_k. Siggers solved (5) for K using Newton-Raphson method[21]. To obtain an initially stable closed loop system, he introduced an artificial large damping to the system[4] for the first few iterations. As for now K can be computed from the eigenvectors of a composite matrix M for much less time[22].

The other difficulty of the state regulator design was the choice of the state weighting matrix Q. It was chosen arbitrarily from guessing until a statisfactory result was obtained. The problem is solved[23] and Q is determined from a dominant eigenvalue shift within controllers' practical limit using the sensitivity coefficients of eigenvalues with respect to the diagonal elements of Q, λ_q.

The composite matrix M[22]

$$M = \begin{bmatrix} A & -S \\ -Q & A' \end{bmatrix} \tag{7}$$

has the following properties:

1. Its eigenvalue distribution on the complex plane is symmetrical to both real and imaginary axes and the eigenvalue vector may be written

$$\lambda = \begin{bmatrix} \lambda_I \\ \lambda_{II} \end{bmatrix} ; \qquad \lambda_I = -\lambda_{II} \tag{8}$$

where λ_I consists of the stable eigenvalues which are the same as those of the closed loop system

$$\dot{y} = Gy \tag{9}$$

where

$$G = A - SK \tag{9a}$$

2. Let the eigenvectors of M be

$$X = \begin{bmatrix} X_I & X_{III} \\ X_{II} & X_{IV} \end{bmatrix} \tag{10}$$

where $[X_I, X_{II}]'$ corresponds to λ_I of M. The Riccati metrix K can be computed from

$$K = X_{II} X_I^{-1} \tag{11}$$

The algorithm of determing Q with dominant eigenvalue shift is as follows[5]

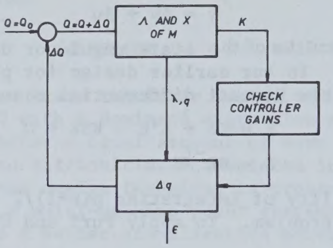


Fig. 1. Algorithm of determining Q

1. Begin with a small Q.
2. Find the eigenvalues and eigenvectors of M
3. Compute K and check controller's practical limit
4. Find the sensitivity coefficient $\lambda,_q$, compute Δq from

   $$\Delta\lambda = \lambda,q \ \Delta q \tag{12}$$

   according to a shift policy stated later, and update Q. Note that the diagonal elements of Q are ordered as vector q for the computation.

The eigenvalues, λ_I, are separated into three groups, the most dominant one, the less dominant ones, and the not dominant ones. The eigenvalue (or pairs) which has the smallest negative real part is the dominant one and always ordered as λ_1, those eigenvalues which have negative real parts up to 5~10 times that of λ_1 are considered as the less dominant ones, and the rest remote eigenvalues as not dominant ones.

The policies for the eigenvalue shift are as follows:

1. Assign a negative real shift ε to λ_1.
2. Keep all the negative real shifts of less dominant eigenvalues within ε, and damp out all the positive real movements.
3. Relax the movements of all the remote eigenvalues. Otherwise it will result in unusual large controller gains.

VARIOUS STATE REGULATOR SCHEMES FOR A ONE MACHINE SYSTEM

In most stabilization studies, a power system may be modelled as one machine connected to an infinite system through a transmission network, Fig. 2. The transmission system may be considered static and the infinite system is assumed to maintain a constant voltage with infinite inertia. Although the synchronous machine can be modelled in great detail because of its many windings, it is found from computed and machine test results that a third order modelling with the field flux linkage ψ_F, the torque angle δ, and the speed ω is sufficient for most stability studies[24]. The complete state equations for the one machine system can be obtained[6] by including an exciter-voltage regulator, Fig. 3, and a governor, Fig. 4.

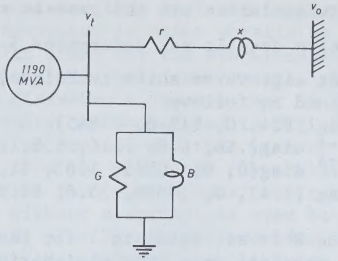

v_t
r
x
v_o
1190 MVA
G
B

Fig. 2 A typical one-machine infinite system

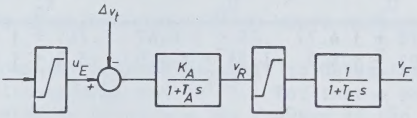

Δv_t
u_E
+
−
$\frac{K_A}{1+T_A s}$
v_R
$\frac{1}{1+T_E s}$
v_F

Fig. 3 A typical exciter-voltage regulator

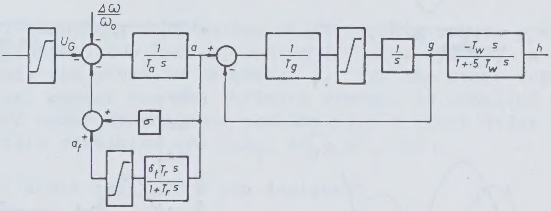

$\frac{\Delta\omega}{\omega_o}$
U_G
−
−
−
$\frac{1}{T_a s}$
a
+
−
$\frac{1}{T_g}$
$\frac{1}{s}$
g
$\frac{-T_w s}{1+.5 T_w s}$
h
+
σ
a_f
+
$\frac{\delta_t T_r s}{1+T_r s}$

Fig. 4 A typical governor hydraulic system

In Fig. 3, U_E is an optimal control signal and Δv_t is the terminal voltage deviation negative feedback. After amplification the signal goes to the exciter with a time constant T_E which can be neglected for solid state exciters. In Fig. 4, U_G is an optimal control signal and $\Delta\omega/\omega_o$ is the per unit speed deviation negative feedback. The lowest block represents the dashpot which usually has a large time constant. The signals a, g, and h represent actuator position, gate opening and waterhead respectively and the last block is called the hydraulic operator.

Several state regulator schemes are investigated:

U_E: With only one optimal excitation control signal U_E. The governor is neglected in this design and the state variables are $(\Delta\psi_F, \Delta\nu_F, \Delta\delta, \Delta\omega)'$.

U_G: The system has both voltage-regulator and governor but with only one optimal governor control signal U_G the state variables are $(\Delta\psi_F, \Delta\nu_F, \Delta\delta, \Delta\omega, a, a_f, g, h)'$.

U_G': Same as U_G except the dashpot is removed. The state variables are $(\Delta\psi_F, \Delta\nu_F, \Delta\delta, \Delta\omega, g, h)'$.

U_E and U'_G: The state variables are the same as the case of U'_G. For data and system equation details see references 5 and 6.

Applying the dominant eigenvalue shift technique, the state weighting matrices, Q's are determined as follows

U_E $Q = \mathrm{diag}[2524, 0, 913.6, 23865]$
U_G $Q = 10^{-4}\,\mathrm{diag}[.56, 4.8, .116, 6.8, .034, .0019, .52, 0]$
U'_G $Q = 10^{-4}\,\mathrm{diag}[0, 0, .0063, 1.83, 31.1, 0]$
U_E and U'_G $Q = \mathrm{diag}\,[1.42, 0, .0859, 25.8, 82.28, .025]$

In the investigations R is set equal to 1 for the first three cases, and to diag [1, 160] for the last case from engineering considerations.

The dominant eigenvalue pairs, which correspond to the mechanical modes of the system, before and after the shift, are as follows

	U_E	U_G	U'_G	$U_E+U'_G$
Before Shift	$.178 \pm j\ 4.77$	$.25 \pm j\ 4.67$	$.715 \pm j\ 4.35$	$.715 \pm j\ 4.35$
After Shift	$-2.07 \pm j\ 4.9$	$-1.35 \pm j\ 4.9$	$-3.7 \pm j\ 4.9$	-4.13 ± 5.33

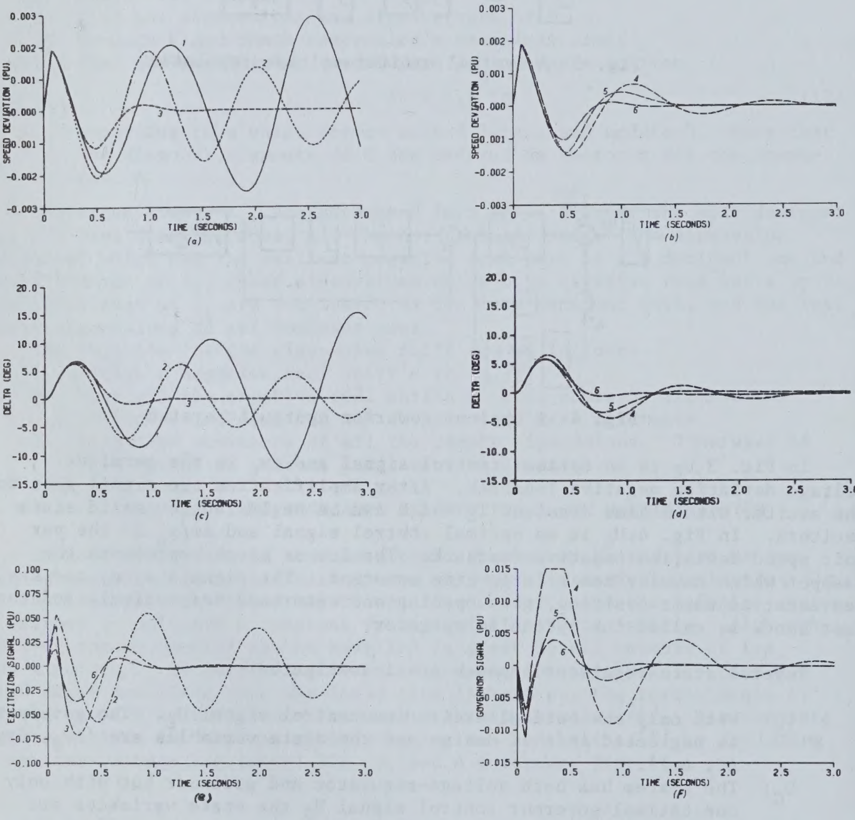


Fig. 5 Nonlinear tests of state regulator design of the one-machine system
1. No control 2. U_{EC} 3. U_E 4. U_G 5. U'_G 6. $U_E+U'_G$

Each linear state regulator thus designed is given a nonlinear test, for which the system is described in great details by high order nonlinear differential equations, 7th order for the synchronous machine alone in addition to controller equations and their constraints. The system dynamics are integrated forward for a disturbance, to see that is the system effectively stabilized. The main results are summarized in Fig. 5, including the case with conventional excitation control and that with no control at all for comparison. It is found from the nonlinear tests that the optimal excitation control U_E is more effective than conventional excitation control U_{EC}, that U'_G control of a governor without a dashpot is even better than U_G control with a dashpot, and that U_E plus U'_G control provides the best results, (Fig. 5 a-d). It also indicates that for the stabilization of the system the conventional excitation control U_{EC} must work harder than the optimal excitation control U_E (Fig. 5e) and the optimal governor control U_G must work much harder than $U_E + U'_G$, (Fig. 5f).

VARIOUS STATE REGULATOR SCHEMES FOR MULTI-MACHINE SYSTEMS

The technique of dominant eigenvalue shift to determine Q[5] can be readily applied to a multi-machine system[6]. For that a good formulation of the multi-machine system is necessary. The feature of the formulation is that while the individual synchronous machine equations have the simplest form in their individual rotating coordinates, they are related to a common static transmission network, and the points of formulation are how to save computation time and storage especially large matrix inversions, and how to keep the information of all parameters for sensitivity studies. The formulation is now available[6].

For the multi-machine stabilization study in this paper, a 4-machine system, i.e., one thermo, two hydros, and an infinite system, is considered. The details and data are given in reference 6. For the state regulator design, each machine, except for the infinite system, is modelled as a 4th order system, a 3rd order synchronous machine plus a first order voltage regulator. The state variables are $(\Delta\psi_F, \Delta v_F, \Delta\delta, \Delta\omega)'$.

Again several state regulators are designed:

U_{EI}: With one optimal excitation control signal of the multi-machine design and applied to machine #3 only.
U_{EM}: With three optimal excitation control signals of the multi-machine design, one for each machine.
U_E: With one optimal excitation control signal designed for machine #3 and with the rest approximated as an infinite bus.
m-U_E: With three optimal excitation control signals of the individual machine design, without considering the interaction of multi-machine dynamics, but applied together to the multi-machine system.

The U_E control of one machine design in the previous section is included for comparison.

Applying the dominant eigenvalue shift technique, the mechanical modes of the system are as listed. When the U_E control of the one machine design applied to the multi-machine system it does not provide much damping to machine #1 or #2 as in the case of U_{EI}. When m U_E controllers are applied together to the multi-machine, although not listed, one of the other two eigenvalues of generator #2 becomes +.169 which causes the instability of the complete system.

Generator	#1	#2	#3
Before Shift	-.013 ± j 7.8	-.018 ± j 7.4	+.177 ± j 3.98
U_{EI} on #3	-1.17 ± j 7.86	-.3 ± j 7.86	-1.88 ± j 3.55
(U_E on #3)	-.084 ± j 7.46	-.1 ± j 7.8	-3.3 ± j 4.5
U_{EM}	-1.01 ± j 7.64	-.448 ± j 7.89	-2.05 ± j 4.04
m-U_E	-4.19 ± j 7.58	-4.63 ± j 7.84	-2.53 ± j 4.47

Like the one machine cases, the linear state regulators thus designed for the multi-machine system are each given a nonlinear test. The main results are summarized in Fig. 6. It is found from the results that, while U_{EI} control of the multi-machine design is more effective than U_E control of the one machine design, for the stabilization of the multi-machine system, U_{EM} control with m controllers of the multi-machine design provides the best answer. On the other hand m-U_E control with m controllers of the individual machine design just do not work, when they are put together for the stabilization of the multi-machine system.

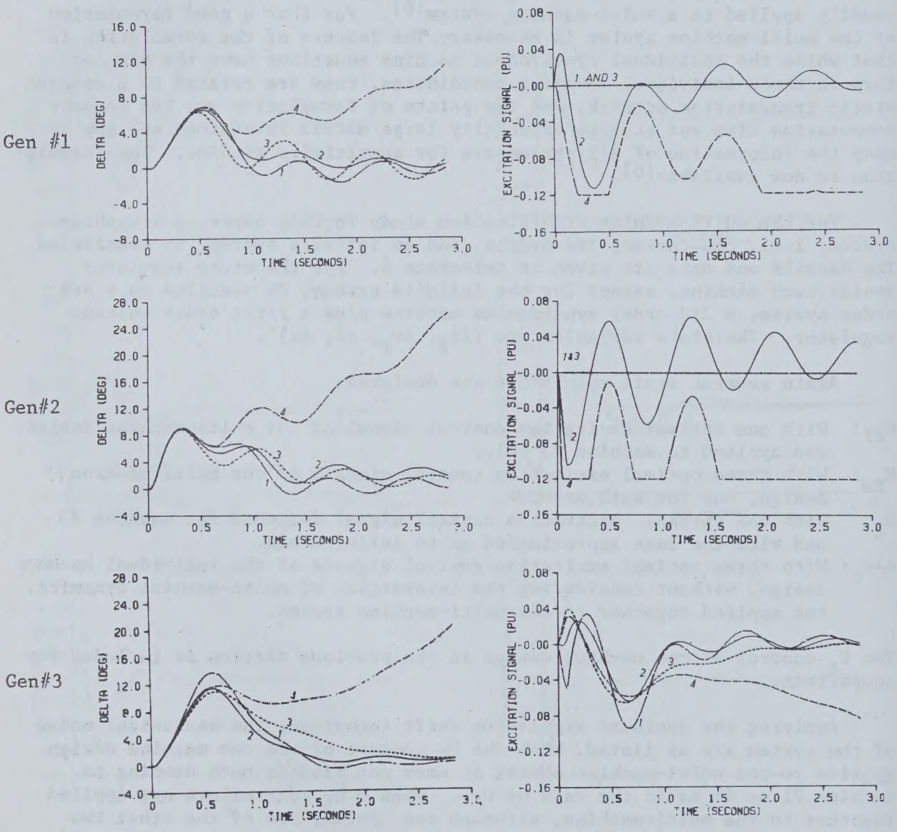


Fig. 6 Nonlinear tests of state regulator design of the multi-machine system

1. U_{EI} 2. U_{EM} 3. U_E 4. m-U_E

CONCLUSION

A new method of determining Q for the state regulator design is presented. For that the dominant eigenvalues are shifted, the movements of the less dominant eigenvalues to the right are damped, and the movements of the remote ones are relaxed. It is also found that if forced movements or shifts are imposed on remote eigenvalues impractical unusually large controller gains will result. The method is general and can be applied to the linear state regulator design of any system.

When this technique is applied to the regulator design of a power system it is found that the optimal control is much more effective than the conventional excitation control and that multi-machine state regulator design gives the best results.

Works are continuing in the fields of wide range power system operating conditions and suboptimal control with an imcomplete set of state variables.

ACKNOWLEDGMENT

The authors wish to acknowledge the financial support of N.R.C. of Canada through the grant A 3626, and the interest shown by their colleagues especially Dr. A.D. Moore, head of the Department of Electrical Engineering at the University of British Columbia.

APPENDIX: SENSITIVITY COEFFICIENT $\lambda_{,q}$

Although Chen and Shen[25] gave two algorithms to compute $\lambda_{,q}$, a new sensitivity formula is developed[5] and expressed in terms of the eigenvector matrix X of equation (10). Since the same eigenvector matrix is required for the computation of K, (11), the computational effort is much reduced.

It is known that the eigenvector matrix of M' can be written in terms of that of M as[22]

$$V = \begin{bmatrix} X_{IV} & X_{II} \\ -X_{III} & -X_{I} \end{bmatrix} \tag{13}$$

Let the i-th eigenvectors X_i and V_i which correspond to a stable eigenvalue λ_i be

$$X_i = (X_{Ii}, X_{IIi})', \tag{14a}$$

and

$$V_i = (X_{IVi}, -X_{IIIi})'. \tag{14b}$$

Since[26]

$$\Delta\lambda_i = \frac{1}{C_i} V_i' \; \Delta M \; X_i \tag{15a}$$

where

$$C_i = V_i' X_i, \tag{15b}$$

and

$$\Delta M = \begin{bmatrix} 0 & 0 \\ -\Delta Q & 0 \end{bmatrix} \tag{16}$$

one has

$$\Delta\lambda_i = \frac{1}{C_i} X_{III}' \quad \Delta Q \; X_{Ii} \tag{17}$$

For the changes in the diagonal elements of Q we write

$$\Delta\lambda_i = \lambda'_{i,q}\ \Delta q \tag{18}$$

where

$$\lambda_{i,q} = (\lambda_{i,q_1}, \lambda_{i,q_2}, \ldots, \lambda_{i,q_n})', \tag{18a}$$

$$\lambda_{i,q_j} = \frac{1}{C_i} X_{IIIi}(j)X_{Ii}(j), \tag{18b}$$

and

$$C_i = \sum_{j=1}^{n} [X_{IVi}(j)X_{Ii}(j) - X_{IIIi}(j)X_{IIi}(j)] \tag{18c}$$

REFERENCES

[1] Yu, Yao-nan, Vongsuriya, K. and Wedman, L.N., "Application of an Optimal Control Theory to a Power System", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-89, pp. 55-62 Jan. 1970.

[2] Vongsuriya, K., "The Application of Lyapunov Function to Power System Stability Analysis and Control", Ph.D. Thesis, Department of Electrical Engineering, the University of British Columbia, Feb. 1968.

[3] Wedman, L.N. and Yu, Yao-nan, "Computation Techniques for the Stabilization and Optimization of High Order Power Systems", 1969 PICA Conference Proceedings.

[4] Yu, Y.N. and Siggers, C, "Stabilization and Optimal Control, Signals for a Power System", Transactions Paper 70TP531-PWR, IEEE Summer Power Meeting, July 1970.

[5] Moussa, H.A.M. and Yu, Yao-nan, "Optimal Power System Stabilization through Excitation and/or Governor Control", Transactions Paper 71 TP 581-PWR, IEEE Power Engineering Society Summer Meeting, July 1971.

[6] Yu, Yao-nan and Moussa, H.A.M., "Optimal Stabilization of a Multi-Machine System" Conference Paper 71 CP 603-PWR, IEEE Power Engineering Society Summer Meeting, July 1971,(accepted for publication).

[7] deMello, F.P. and Concordia, C., "Concepts of Synchronous Machine Stability as Affected by Excitation Control", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-88, pp. 316-329, April 1969.

[8] Schleif, F.R., Hunkins, H.D., Martin, G.E. and Hattan, E.E., "Excitation Control to Improve Powerline Stability", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-87, pp. 1426-1434, June 1968.

[9] Ellis, H.M., Hardy, J.E., Blythe, A.L. and Skooglund, J.W., "Dynamic Stability of the Peace River System", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-85, pp. 586-600, June 1966.

[10] Hanson, O.W., Goodwin, C.J. and Dandeno, P.L., "Influence of Excitation and Speed Control Parameters in Stabilizing Intersystem Oscillations", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-87, pp. 1306-1313, May 1968.

[11] Byerly, R.T., Keay, F.W. and Skooglund J.W., "Damping of Power Oscilla-

tions in Salient Pole Machines with Static Exciters", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-89, pp. 1009-1021, July/August 1970.

[12] Fosha Jr., C.E. and Elgerd, O.I., "The Megawatt-Frequency Control Problem: A new Approach via Optimal Control Theory", IEEE Transactions, Vol. PAS-89, pp. 563-577, April 1970.

[13] Mittelstadt, W.A., and Gorger, D.E., "Application of Optimal Control Theory to Power System Stability Studies", Third Southeastern Symposium on System Theory, April 1971, Georgia Institute of Technology.

[14] Davison, E.J. and Rau, N.S., "The Optimal Feedback Control of a Synchronous Machine", Transactions Paper 71 TP 102-PWR, IEEE Power Engineering Society Winter Meeting, Feb. 1971.

[15] Kelly, D.H. and Rahim A., "Closed-loop Optimal Excitation Control for Power System Stability" Transactions Paper 71 TP 103-PWR, IEEE Power Engineering Society Winter Meeting, Feb. 1971.

[16] Miniesy, S.M. and Bohn, E.V., "Two-level Control of Interconnected Power Plants", Transactions Paper 71 TP 624-PWR, IEEE Power Engineering Society Summer Meeting, July 1971.

[17] Iyer, S.N. and Cory, B.J., "Optimization of Turbo-Generator Transient Performance by Differential Dynamic Programming", Transactions Paper 71 TP 72-PWR, IEEE Power Engineering Society Winter Meeting, Feb. 1971.

[18] Anderson, J.H., "The Control of a Synchronous Machine Using Optimal Control Theory", Proceedings of the IEEE, Vol. 59, pp. 25-35, Jan. 1971.

[19] Puri, N.N., and Gruver, W.A., "Optimal Control Design via Successive Approximations", Preprints of the 1967 Joint Automatic Control Conference, pp. 335-343, University of Pennsylvania.

[20] Davison, E.J. and Man, F.T., "The Numerical Solution of A'Q + QA = -C", IEEE Transactions, Vol. AC-13, pp. 448-449, 1968.

[21] Blackburn, T.R., "Solution of the Algebraic Matrix Ricatti Equation via Newton-Raphson Iteration; Presented at the JACC Meeting in 1968.

[22] Potter, J.E., "Matrix Quadratic Solutions", SIAM, Journal of Applied Math, Vol. 14, No. 3, pp. 496-506, May 1966.

[23] Moussa, H.A.M., "Linear Optimal Stabilization and Representation of Multi-Machine Power Systems", Ph.D. Dissertation, University of British Columbia, July 1971.

[24] Yu, Yao-nan, et al., "Optimal Stabilization and Resynchronization Studies on a Dynamic Power System Model", Presented at 1970 Annual General Meeting, E.I.C.

[25] Chen, R.T. and Shen, D.W., "Sensitivity Analysis and Design of Multivariable Regulators using a Quadratic Performance Criterion", Preprints of the 1968 JACC Meeting.

[26] Faddeev, D.K. and Faddeeva, V.N., "Computational Methods of Linear Algebra", Freeman, 1963, (book).

DESIGNING SHORT-WORD RECURSIVE DIGITAL FILTERS*

K. STEIGLITZ
Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08540

ABSTRACT

This paper deals with the problem of designing recursive digital filters with limited word length coefficients to meet magnitude specifications in the frequency domain. Schuessler and Avenhaus have reported results on this problem obtained by rounding the coefficients of an optimal filter and applying a modified univariate search. We describe here another search method, based on the Pattern Search of Hooke and Jeeves, and relate some computational experience with the problem.

I. INTRODUCTION

Schuessler and Avenhaus [1-3] have reported results on the problem of designing recursive digital filters with limited word length coefficients to meet magnitude specifications in the frequency domain. They begin with a filter optimal in the minimax sense, round its coefficients, and apply a discrete optimization method to obtain a short-word filter which meets specifications. Their optimization algorithm consists of univariate search with two important modifications: (1) a backtracking feature which returns, after a local optimum has been found, to previous next-best points in parameter space; and (2) a more powerful local search which entails generating a four-dimensional subspace determined by the most sensitive parameters.

We shall describe here another algorithm, based on the well-known Pattern Search of Hooke and Jeeves [5], which yields similar results, and has the advantages of being easy to program and flexible in trading off computing time with quality of solutions. The Pattern Search is a univariate search with an acceleration feature, that has been modified in two ways: (1) the search is randomized so that a variety of locally optimal solutions can be generated; and (2) the neighborhood ex-

---

* This work was supported by NSF Grant GJ-965, and U.S. Army Research Office - Durham under contract DAHCO4-69-C-0012.

plored at each point can be expanded, as in the Schuessler-Avenhaus method.

One purpose of this paper is to show that the problem is amenable to the fairly standard optimization algorithm of Hooke and Jeeves. Another purpose is to demonstrate that for a given problem a wide variety of locally optimum short-word filters exist, and that a randomized procedure for producing an assortment of solutions seems useful. In this respect the problem is similar to other combinatorial problems, such as the traveling salesman problem [8] and the problem of designing low cost redundant communication networks [9].

II. STATEMENT OF THE PROBLEM

We consider the cascade form transfer function

$$Y(\omega) = \prod_{k=1}^{K} \frac{1 + a_k z^{-1} + z^{-2}}{1 + b_k z^{-1} + c_k z^{-2}} \qquad (1)$$

where (1) $z = e^{j\pi\omega}$, ω is the frequency in fractions of the Nyquist frequency;

(2) $|a_k| < 2$, so that the zeros are complex and on the unit circle;

(3) $c_k < 1$, $1 + c_k + b_k > 0$, $1 + c_k - b_k > 0$, so that the poles are inside the unit circle;

(4) $2^Q a_k$, $2^Q b_k$, and $2^Q c_k$ are integers.

From (2), (3) and (4) we can conclude that a_k and b_k require $Q + 2$ bits (a sign bit, a unit bit, and Q fraction bits); and that c_k requires $Q+1$ bits (a sign bit, and Q fraction bits). The coefficient quantization step-size is 2^{-Q}.

We are given an ideal desired magnitude characteristic Y^d. Define the error function

$$e(\omega) = \frac{1}{\delta(\omega)} \left| A|Y(\omega)| - Y^d \right|, \quad \omega \in W \qquad (2)$$

where A is a positive gain constant, $W \subset [0,\pi]$ is a closed set of frequencies of interest, and $\delta(\omega)$ is a tolerance function. The problem is to choose the coefficients a_k, b_k, and c_k so that

$$F(a_k, b_k, c_k) = \min_{A} \max_{\omega \in W} e(\omega) \qquad (3)$$

is as small as possible. In particular, whenever $F \le 1$, we say we have a design which meets the tolerance specifications $\delta(\omega)$.

Two remarks: First, the gain A is not constrained in precision, since a wide range of overall gain is usually tolerable. Second, the zeros are restricted to the unit circle since this reduces the complexity of the filter and is a good assumption for filters with bands of desired rejection, such as lowpass, highpass, bandpass, etc. This assumption can be relaxed for other cases with only trivial changes in what follows.

III. HOOKE AND JEEVES' PATTERN SEARCH [5]

Figures 1 and 2 show flowcharts of the basic pattern search, and Table 1 summarizes the variables used. The flowcharts are of a specialized type called D-charts [6], and have no explicit go-to control statements. Two control statements are used: an if (condition) then A else B, indicated by a simple branching; and a do A while (condition), indicated by an iteration node "⊗" with the condition on a branch to the right. The first time the condition is violated at the node, execution continues downward. All variables are global.

The basic strategy of the main program (MAIN in Fig. 1) is the following: the current value of the parameter vector is Ψ, and φ is the exploratory value. Subroutine EXPLOR (Fig. 2) performs a univariate search by incrementing and decrementing each component of φ, accepting any improvements in the function F. If an improvement is found, θ is set to Ψ, saving the old value; Ψ is set to φ, updating the current value; and φ is set to $2\varphi-\theta$, thereby attempting to extrapolate the pattern of the successful move. EXPLOR is then invoked for a local exploration about φ, and the process continued until no improvements are found. If at least one improvement was found during this process ($S>0$), the strategy is applied again in an attempt to build up a new pattern. When finally no improvements are found in this attempt, ($S=0$), the stepsize Δ is reduced by a factor of 2 and the process repeated. The algorithm terminates when Δ falls below some prescribed value δ.

In order that condition (4) above hold throughout, it is sufficient that it hold for the initial value of the parameter vector, that Δ_o be an integer power of 2 no less than 2^{-Q}, and that δ be at least 2^{-Q}. Generally, it was found effective to take $\Delta_o = 2^{-Q+1}$ or 2^{-Q}, and $\delta = 2^{-Q}$. Too large a value of Δ_o tends to encourage prematurely large steps early in the algorithm, which leaves the parameter vector trapped at poor local optima.

IV. RANDOMIZATION

The original Pattern Search of Hooke and Jeeves enumerates the coordinates in the natural order. The course of the algorithm is sensitive to the order in which the coordinates are explored, however, since improvements are retained whenever encountered. It was found that using a random ordering of coordinates produced a variety of final local optima, the best of which is selected as the final solution. In this way, more computer time can be invested in the hope of improving the solution. The set $\{1,...,N\}$ is randomly shuffled by using a pseudorandom number generator with different initial seeds, and an algorithm described by Knuth [10].

V. TWO-OPT

As reported by Avenhaus [3], the univariate search is rather weak, and often leads to local optima which are relatively poor. For this reason, a stronger version of EXPLOR was developed, using the routine TWO-OPT shown in a dotted box in EXPLOR. Fig. 3 shows a D-chart for TWO-OPT, which works as follows: if no improvement is found by changing coordinate i = LIST (k), two coordinates at a time are changed: i and j = LIST (ℓ), $\ell > k$. In this way, stronger local optima are found, at the expense of increasing the size of the neighborhood search. This approach is similar to the use of 3-opt instead of 2-opt for the traveling salesman problem [8].

VI. COMPUTATIONAL RESULTS

(The values of F given are taken from a finite grid of frequency points, and may be slightly optimistic.)

Example 1 This is the bandpass example of Schuessler and Avenhaus [1-3]:

$$K = 4,\ N = 12,\ W = W_p \cup W_s, \text{ where}$$

W_p = [.411111, .466666] is the passband, and W_s = [0.,.383333] $\cup$ [.494444,1.] is the stopband.

$$Y^d = 1 \quad \omega \in W_p$$

$$\quad\ 0 \quad \omega \in W_s$$

$$\delta(\omega) = .03 \quad \omega \in W_p$$

$$\qquad\quad .01 \quad \omega \in W_s$$

The starting point was obtained by rounding the coefficients of the elliptic filter used by Avenhaus [4]. The complete algorithm with TWO-OPT was run 3 times with 3 different initializations of the random number generator, $\Delta_o = 2^{-5}$ and $\delta = 2^{-6}$. An unacceptable solution with $F = 1.80$ was produced once, and twice the following acceptable local optimum with $F = .802$ was found (in hexadecimal):

a	b	c
.70	-.48	.E8
-1.18	-.74	.E8
- .04	-.30	.F8
- .BC	-.90	.F8

This differs from Avenhaus' solution [4] with $F = .732$ by 2^{-6} in a_2, and this discrepancy is due to the fact that a finite and fixed grid of frequency points was used, rather than a search for minima and maxima of F. Each run required between 1788 and 2650 evaluations of F.

Other runs, without TWO-OPT, and with various Δ_o, all failed to produce acceptable solutions, thus showing the necessity of using the more powerful version of EXPLOR.

Example 2 This is another example due to Schuessler and Avenhaus [2,3], and data provided by Avenhaus [4] provided the lowpass elliptic filter rounded to obtain a starting point.

$$K = 5,\ N = 15,\ W = W_p \cup W_s, \text{ where}$$

$$W_p = [0.,.333333] \text{ and } W_s = [.376111,1.]$$

$$Y^d = 1 \quad \omega \in W_p$$

$$\quad\ 0 \quad \omega \in W_s$$

$$\delta(\omega) = .01 \quad \omega \in W_p$$

$$\qquad\quad .001 \quad \omega \in W_s$$

Five different random starts with $\Delta_o = \delta = 2^{-7}$, produced 5 distinct solutions with F from .629 to 1.040, using between 1158 and 2302 evaluations of F. The best solution is not quite as good as Avenhaus' solution [4] with F = .581, but is acceptable. In hexadecimal it is

a	b	c
-.BE	-1.12C	.60
-.AA	-1.11A	.8C
-.60	-1.104	.BC
.52	- .F6	.E0
1.A	- .F4	.F6

Three other runs with $\Delta_o = 2^{-6}$, $\delta = 2^{-7}$ produced similar results, with F from .752 to .917. EXPLOR without TWO-OPT also produced some acceptable local optima, with the best of 4 local optima having F = .845, and using 356 evaluations of F.

Example 3 This the elliptic lowpass filter given by Gold and Rader [7]:

$K = 4$, $N = 12$, $W = W_p \cup W_s$, where

$W_p = [0.,.4]$ and $W_s = [.44,1.]$

$$Y^d = 1 \quad \omega \in W_p$$

$$0 \quad \omega \in W_s$$

$$\delta(\omega) = .03 \quad \omega \in W_p$$

$$.001 \quad \omega \in W_s$$

The margin for improvement here is not great, since the infinite precision filter has $F \cong .877$. One run with $\Delta_o = 2^{-8}$, $\delta = 2^{-9}$ produced a local optimum with F = .935. Three runs with $\Delta_o = 2^{-7}$, $\delta = 2^{-8}$ produced a best solution with F = 1.189. Two runs with $\Delta_o = \delta = 2^{-8}$ produced a best solution with F = 1.080. Two runs with $\Delta_o = \delta = 2^{-7}$ produced a best solution with F = 1.018. Notice that this best 7-bit filter is slightly better than the best 8-bit filter, an anomoly due to the suboptimal nature of the solutions. The best 7-bit filter is

a	b	c
1.1A	-1.12C	.6C
.5A	- .E2	.AA

a	b	c
-.32	-.AA	.DA
-.5C	-.98	.F6

CONCLUSION

It is an open question whether a practical and exact algorithm can be found to find optimal solutions to this problem. On the other hand, heuristically motivated search techniques, such as the Schuessler-Avenhaus algorithm and the one described here, provide practical solutions with reasonable computing costs.

ACKNOWLEDGEMENTS

The author is indebted to Prof. Dr. W. Schuessler of the University Erlangen-Nuernberg and Dr. J.F. Kaiser of Bell Telephone Laboratories for helpful discussions about this problem, and to Dipl.-Ing. E. Avenhaus for kindly supplying the author with data.

REFERENCES

1. E. Avenhaus and W. Schuessler, "On the Approximation Problem in the Design of Digital Filters with Limited Wordlength," Archiv Der Elektrischen Vebertragung, Vol. 24, No. 12, pp. 571-572; December 1970.

2. W. Schuessler, "On the Approximation Problem in the Design of Digital Filters," Proc. Fifth Annual Princeton Conference on Information Sciences and Systems, Princeton, N.J., pp. 54-63; March 25-26, 1971.

3. E. Avenhaus, "An Optimization Procedure to Minimize the Wordlength of Digital Filter Coefficients," 1971 Imperial College Symposium on Digital Filtering, Imperial College, London; August 31, September 1,2, 1971.

4. E. Avenhaus, private communication.

5. R. Hooke and T.A. Jeeves, "'Direct Search' Solution of Numerical and Statistical Problems," JACM, Vol. 8, No. 2, pp. 212-229; April 1961.

6. J. Bruno and K. Steiglitz, "The Expression of Algorithms by Charts," to appear in CACM.

7. B. Gold and C.M. Rader, Digital Processing of Signals, McGraw-Hill, 1969. See Chapter 3, Design Example 1, pp. 86-90.

8. S. Lin, "Computer Solutions of the Traveling Salesman Problem," BSTJ, pp. 2245-2269; December 1965.

9. K. Steiglitz, P. Weiner, and D.J. Kleitman, "The Design of Minimum-Cost Survavable Networks," IEEE Trans. on Circuit Theory, Vol. CT-16, No. 4, pp. 455-460; Nov. 1969.

10. D.E. Knuth, The Art of Computer Programming, Vol. 2, "Seminumerical Algorithms," Addison-Wesley, 1969. See Section 3.4.2, Algorithm P, p. 125.

Table 1 Description of Variables

N	dimension of parameter space
Δ	step-size
Δ_o	initial step-size
δ	minimum step-size (for terminating search)
Ψ	current value of parameter vector
φ	exploratory value of parameter vector
θ	previous value of parameter vector
F( )	function to be minimized, see Eq. (3)
FΨ	F(Ψ)
Fφ	F(φ)
S	current number of consecutive successful steps
FNEW	F( ) at exploratory point
LIST( )	pseudorandom permutation of {1,...,N}
k,ℓ	counting indexes
i,j	dimension indexes
SAVE, SAVEi, SAVEj	temporary variables to store old values of parameters

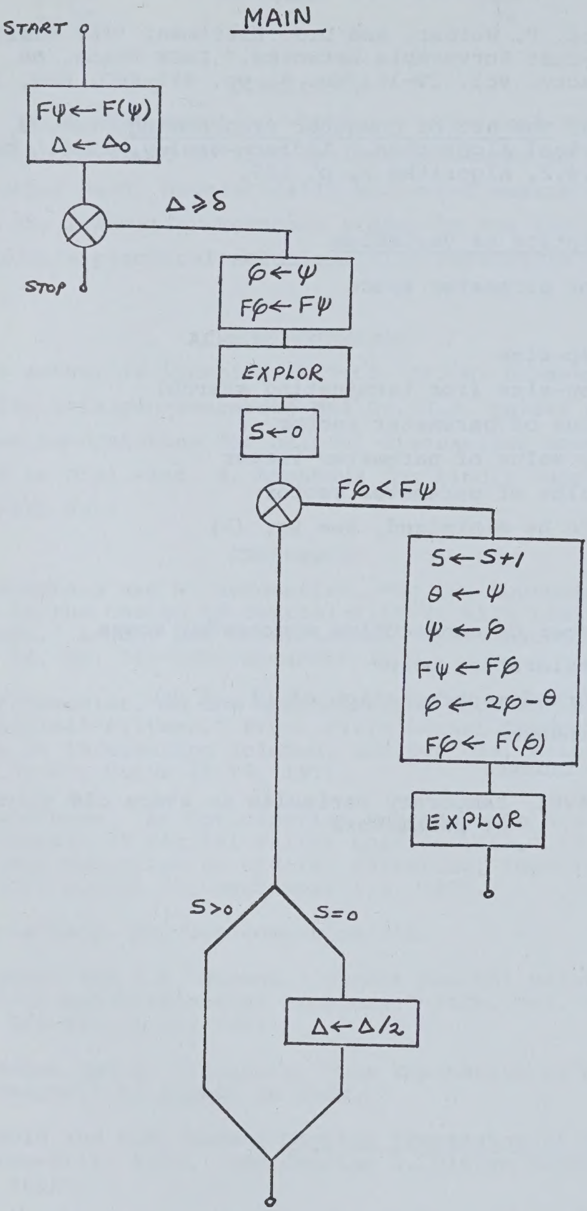

MAIN
START
Fψ ← F(ψ)
Δ ← Δ0
Δ ≥ δ
STOP
φ ← ψ
Fφ ← Fψ
EXPLOR
S ← 0
Fφ < Fψ
S ← S+1
θ ← ψ
ψ ← φ
Fψ ← Fφ
φ ← 2φ−θ
Fφ ← F(φ)
EXPLOR
S>0
S=0
Δ ← Δ/2

Figure 1

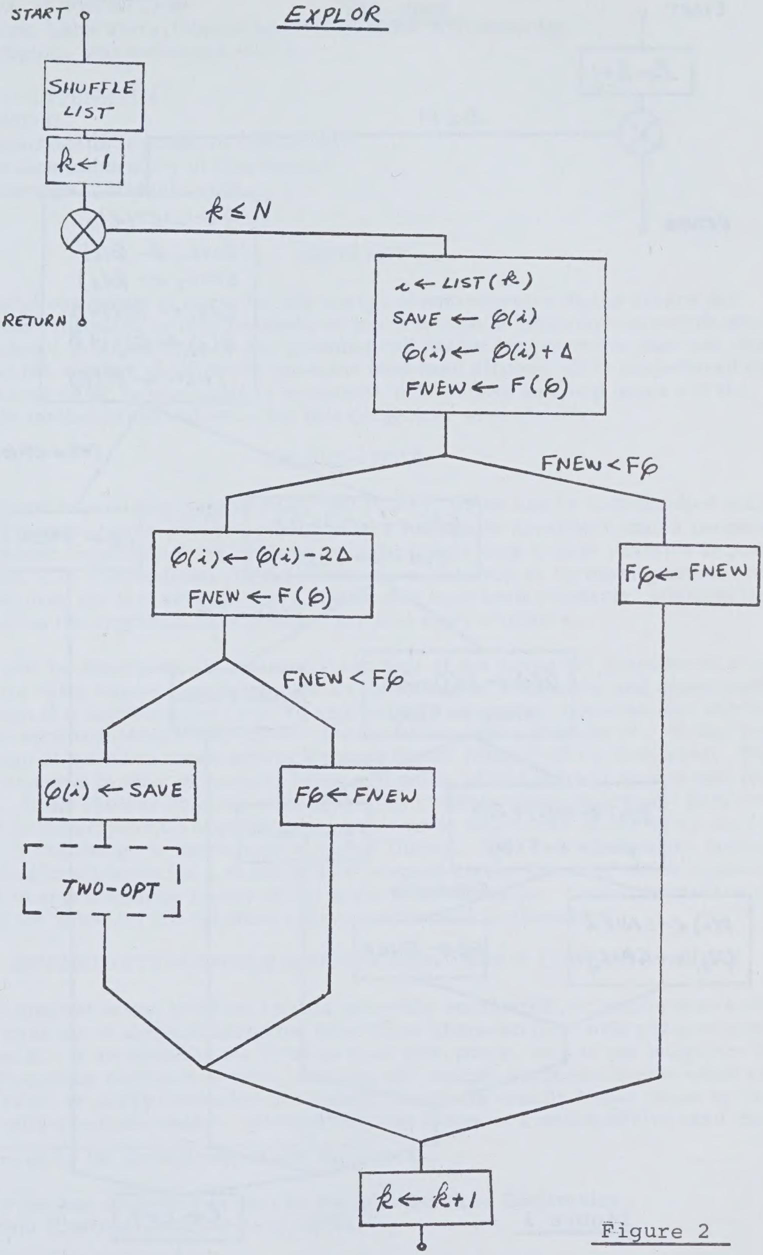

EXPLOR
START
SHUFFLE LIST
$k \leftarrow 1$
$k \leq N$
RETURN
$i \leftarrow LIST(k)$
SAVE $\leftarrow \phi(i)$
$\phi(i) \leftarrow \phi(i) + \Delta$
FNEW $\leftarrow F(\phi)$
FNEW < Fϕ
$\phi(i) \leftarrow \phi(i) - 2\Delta$
FNEW $\leftarrow F(\phi)$
Fϕ $\leftarrow$ FNEW
FNEW < Fϕ
$\phi(i) \leftarrow$ SAVE
Fϕ $\leftarrow$ FNEW
TWO-OPT
$k \leftarrow k+1$

Figure 2

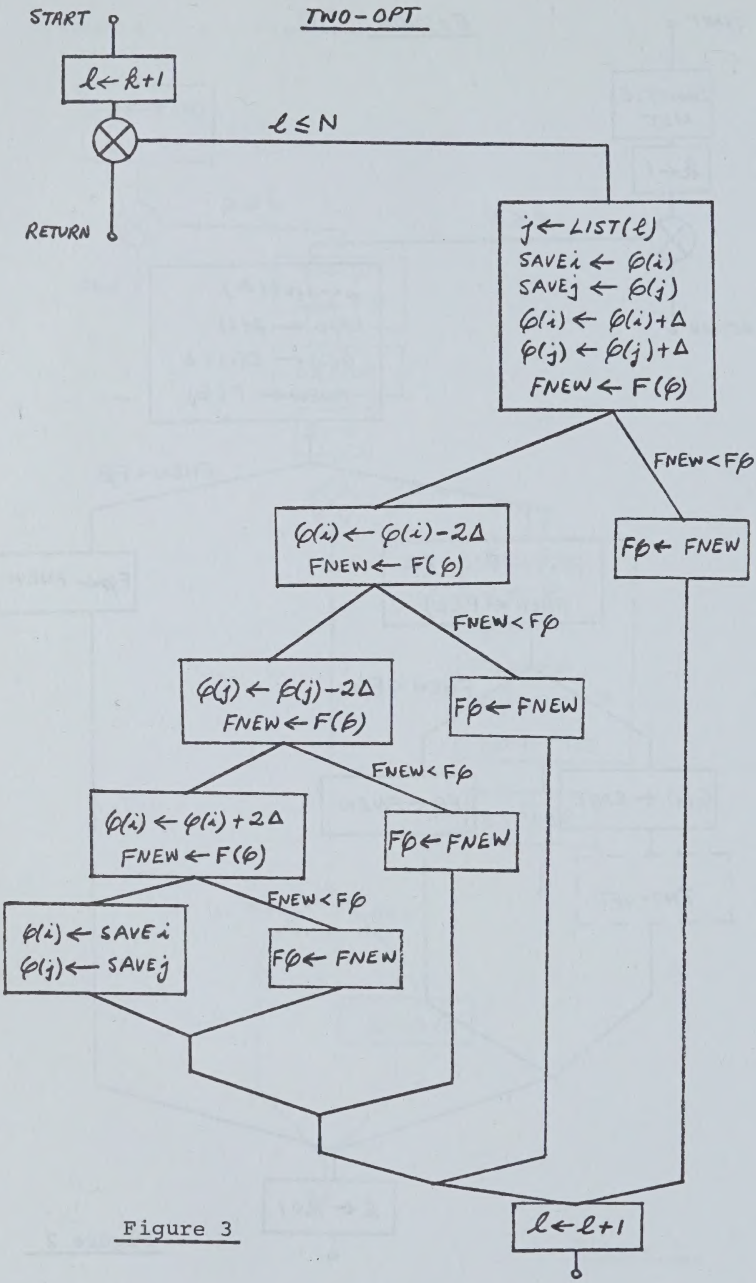

START
TWO-OPT
ℓ←k+1
ℓ≤N
RETURN
j←LIST(ℓ)
SAVEi ← φ(i)
SAVEj ← φ(j)
φ(i) ← φ(i)+Δ
φ(j) ← φ(j)+Δ
FNEW ← F(φ)
FNEW<Fφ
Fφ ← FNEW
φ(i) ← φ(i)−2Δ
FNEW ← F(φ)
FNEW<Fφ
Fφ ← FNEW
φ(j) ← φ(j)−2Δ
FNEW ← F(φ)
FNEW<Fφ
Fφ ← FNEW
φ(i) ← φ(i)+2Δ
FNEW ← F(φ)
FNEW<Fφ
φ(i) ← SAVEi
φ(j) ← SAVEj
Fφ ← FNEW
ℓ←ℓ+1

Figure 3

ON OPTIMUM NONRECURSIVE DIGITAL FILTERS

E. M. HOFSTETTER*
Lincoln Laboratory, Massachusetts Institute of Technology
Lexington, Massachusetts 02173

A. V. OPPENHEIM+
J. SIEGEL
Massachusetts Institute of Technology
Research Laboratory of Electronics
Cambridge, Massachusetts

ABSTRACT

Several optimality criteria for the design of nonrecursive digital filters are discussed. Attention is then focussed on the criterion of minimum transition bandwidth subject to constraints on the maximum allowable errors in the pass and stop band and the number of allowable pass and stop band ripples. It is conjectured that the optimum filter in this class is equiripple in both pass and stop bands and the available mathematical evidence for this conjecture is sketched.

INTRODUCTION

With the exception of tapped delay line filters, which can be implemented exactly as nonrecursive digital filters, analog filters with finite duration impulse response are of minor importance. In contrast, digital filters with a finite duration impulse response, i.e. nonrecursive digital filters have received an increasing amount of attention over the last several years. With this increased emphasis, attention has focussed on the approximation problem for this class of filters.

As will be described, the squared magnitude of the frequency response of a nonrecursive filter can be represented as a trigonometric polynomial and consequently the design of a nonrecursive filter to approximate a desired characteristic corresponds to approximating the specified characteristic with a polynomial. In this paper the design of optimum nonrecursive lowpass digital filters will be discussed. There are of course a number of ways in which optimality can be defined for lowpass filter design. Some choices coincide with optimality criteria which classically have been applied to the polynomial approximation problem in which case such theory can be applied to the design of nonrecursive digital filters. This paper considers primarily optimal approximation for a nonrecursive lowpass digital filter for which classical results in approximation theory do not seem to be available. Some conjectures regarding the nature of the optimum approximation will be discussed.

THE APPROXIMATION PROBLEM FOR NONRECURSIVE LOWPASS FILTERS

The approximation problem as it is generally considered for nonrecursive digital filters consists of approximating the ideal filter characteristic with a trigonometric polynomial. If we consider the filter to have zero phase, so that the imaginary part of the frequency response is zero, then the unit sample response is even about $k=0$. Such a filter is easily converted to a causal filter with exactly linear phase by cascading with a suitable delay. Alternatively the design of a nonrecursive zero phase

---

*Sponsored by the Department of the Air Force.

+This work was supported in part by the Joint Services Electronics Program (Contract DAAB07-71-C-0300.)

filter can be converted to the design of a minimum phase filter.(1) Thus when we consider the approximation with a zero phase characteristic we may either utilize this design for a linear phase filter or a minimum phase filter. For convenience, however, the discussion will be phrased in terms corresponding to the design of a linear phase filter.

Since the unit sample response h(k) is even, i.e. $h(k) = h(-k)$ the frequency response can be expressed as a cosine series of the form

$$G(F) = H(e^{j2\pi F}) = h(0) + 2\sum_{k=1}^{n} h(k)\cos k\,2\pi F \tag{1}$$

where the total duration of the unit sample response is $2n + 1$. Equivalently Eq. (1) can be expressed in terms of trigonometric polynomial of the form

$$G(F) = P\,(\cos 2\pi F)$$

$$\text{where } P(x) = \sum_{k=0}^{n} c_k x^k \tag{2}$$

Because of the form of eq. (1) or (2) the frequency response is symmetric about $F=1/2$ and consequently we need only consider a specification of the filter for F in the interval 0 to 1/2. Then according to eq. (2), we can equivalently approximate the desired characteristic with a polynomial P(x). By mapping the interval $-1\le x\le 1$ to the interval $0 \le F \le 1/2$ according to $x = \cos 2\pi F$, the polynomial approximation is mapped to the filter frequency response. Figure 1 illustrates an acceptable lowpass frequency characteristic versus F and Figure 2 illustrates the same filter as a function of the variable x. It should be noted that in figure 1 the frequency response has zero slope at $F=0$ and $F=1/2$ whereas in Fig. 2 it does not. From the form of eq. (2) it should be clear that whether or not the polynomial P(x) has zero slope at $x= -1$ and $x=1$, the trigonometric polynomial $P\,(\cos 2\pi F)$ always has zero slope at $F=0$ and $F=1/2$.

The parameters typically of interest in characterizing a lowpass filter are the filter order, the passband and stopband edges F_p and F_s, the transition bandwidth $|F_p - F_s|$, the passband ripple δ_p and the stopband ripple δ_s. Clearly it would be desirable to design the filter with the smallest order, transition bandwidth and passband and stopband ripple. Since there are tradeoffs between these parameters, a particular optimization strategy can only be directed toward one parameter. One approach to the optimization is based on the use of linear programming as pursued by Helms(2) and Rabiner(3). In that approach the filter order, and the parameters F_p, F_s and δ_p are specified and linear programming is used to minimize the stopband ripple. In an approach pursued by Parks(4), the filter order, the passband and stopband edges, and the ratio of δ_p to δ_s is fixed. An exchange algorithm is then applied to minimizing δ_p and consequently also δ_s subject to the constraint that the ratio of passband to stopband ripple remain constant. A third approach is directed toward a minimization of the transition bandwidth. This approach is suggested in part by the importance of transition bandwidth in filter design but primarily by the availability of design algorithms(5, 6, 7) that lead to filters that appear to be optimum according to this criterion. In particular, there are several related design techniques for which the design specifications are the number of passband ripples n_p, the number of stopband ripples n_s and the size of the passband and stopband ripples δ_p, and δ_s respectively. These design algorithms result in filters for which the characteristics in the passband are equiripple and the characteristics in the stopband are equiripple. The remainder of this paper is directed toward a discussion of the relationship between the optimality criterion of minimizing the transition bandwidth and these equiripple designs.

MINIMUM TRANSITION BANDWIDTH FILTERS

The first class of filters to be considered consists of those for which the polynomial P(x) has exactly n_p ripples in the passband $x_p \le x \le 1$ and exactly n_s ripples in the stopband $-1 \le x \le x_s^p$ and an order $n = n_p + n_s + 1$. A ripple is defined to be a zero of P'(x) counting multiplicity. Furthermore, we require that within the stop band $| P(x)| \le \delta_s$ and within the passband $| P(1) - 1| \le \delta_p$. Filters within this class are to be compared with regard to their transition bandwidth $| F_p - F_s|$ where $x_p = \cos 2\pi/F$, and $x_s = \cos 2\pi F_s$. It is not hard to show that there is at least one polynomial P*(x) in the class for which the transition bandwidth is smallest, i.e. for which $| F_p - F_s|$ is equal to the greatest lower bound of all the transition bandwidths associated with filters in the class. It is conjectured that P*(x) has the following characteristics:

Conjecture:

P* alternatingly achieves the values $1 \pm \delta_p$ exactly $n_p + 2$ times in the passband and alternatingly achieves the values $\pm \delta_s$ exactly $n_s + 2$ times in the stopband. This type of behavior (which we will refer to as maximally equiripple behavior) is illustrated in Fig. 3 for the case $n_s = 2$, $n_p = 3$.

Discussion: We have not succeeded in proving this conjecture in the full generality stated above; however, a proof can be given if the conjecture is weakened by making the additional assumption that all the ripples of P*(x) are simple (of multiplicity one). This in fact is the class of filters generated by the algorithms in references (5, 6, 7). The nature of the argument will be sketched but a formal proof is omitted.

First we note that all the ripples of P*(x) must lie in the interior of the passband and stopband and that $P^*(-1) = \pm \delta_s$ and $P^*(1) = 1 \pm \delta_p$. If this were not the case, then P*(x) could be subjected to a linear compression in x of the form $x' = a x + b$, $0 < a < 1$ resulting in a new polynomial P*(ax+b) that meets the above requirements and has a smaller transition bandwidth than P*(x). Since P*(x) has the minimum possible transition bandwidth, this is impossible and the above statement must be true.

Now let us denote the locations of the passband ripples by x_{pi}, $i = 1, \ldots n_p$ and the locations of the stopband ripples by x_{si}, $i = 1, \ldots n_s$. If $P^*(x_{pi}) = 1 \pm \delta_p$, $i=1, \ldots n_p$ and $P^*(x_{si}) = \pm\delta_s$, $i = 1, \ldots n_s$, then the conjecture is true because it is easy to verify that P*(x) must exhibit the desired alternating behavior at the $n_p + 2$ passband points $x_p, x_p, \ldots n_{pn_p}$, 1 and the $n_s + 2$ stopband points $-1, x_{s1}, \ldots x_{sn_s}\ x_s$. Therefore, in order for the conjecture to be false, there must exist at least one ripple x_{si} or x_{pi} for which $P^*(x_{si}) \ne \pm \delta_s$ or $P^*(x_{pi}) \ne 1 \pm \delta_p$. Assume that this is the case as is illustrated in Fig. 4. It will now be argued that P*(x) can be modified in such a way as to reduce its transition bandwidth which contradicts the optimality of P*(x). This contradiction establishes the truth of the modified conjecture.

This modification procedure is illustrated in Fig. 4 in which a stopband ripple has been assumed not to achieve the tolerance limit $\pm \delta_s$. It is now possible to draw a line such as aa' that lies in the tolerance band but does not intersect this ripple. The polynomial q(x) is now constructed as shown in Fig. 4 and the modified polynomial $P^*(x) + \lambda\, q(x)$, $\lambda > 0$ is formed. Note that the order of q(x) will not be greater than n so that the modified polynomial $P^*(x) + \lambda\, q(x)$ has the correct order. Furthermore, since the ripples of q(x) are out of phase with all the ripples of P*(x) except the one that did not intersect the line aa', it is obvious that, as long as λ is made sufficiently small, $P^*(x) + \lambda\, q(x)$ will meet the tolerance specifications in $(-1, x_s)$ and $(x_p, 1)$. In addition it is easy to verify that choosing λ small enough leads to the conclusion that $P^*(x) + \lambda\, q(x)$ like P*(x) itself, has n_s simple ripples in the stopband and n_p simple ripples in the passband. But, since $q(x_s) < 0$ and $q(x_p) > 0$, the stopband edge of $P^*(x) + \lambda\, q(x)$ is greater than x_s and the passband edge of $P^*(x) + \lambda\, q(x)$ less than x_p. This means that the transition bandwidth of $P^*(x) + \lambda\, q(x)$ is smaller than that of $P^*(x)$ and this contradiction establishes the weakened form of the conjecture.

The argument sketched above has not been used to prove the conjecture in its full generality because, in this more general case, no way has been found to show that the modified polynomial $P^*(x) + \lambda\, q(x)$ has exactly n_p ripples in the passband and n_s ripples in the stopband.

In the class of filters discussed above the filter characteristic was constrained to have the maximum allowable number of ripples for real x between -1 and 1. We next consider the class of polynomials defined exactly as above with the exception that now the polynomials are only constrained to have no more than n_s ripples in the stopband and no more than n_p ripples in the passband. At first glance, it might appear that, even in this modified class of polynomials, the polynomial with the minimum transition bandwidth is again a maximally equiripple polynomial. This is not true however as demonstrated in the examples prescribed in Figs. 5 and 6. Fig. 5 depicts a maximally equiripple filter response. Fig. 6 shows a filter response with one ripple missing in the stopband. The transition bandwidth of the first filter is $|F_p - F_s| = 0.13$ and that of the second filter $|F_p - F_s| = 0.115$. The latter transition bandwidth is smaller and so we must conclude that the optimum filter in this new class of filters does not necessarily have the maximum allowable number of ripples.

In the example shown in Fig. 6 the ripple that is mssing from the stopband is located just to the right of the passband at x = 1. This ripple is not obvious on a plot of G(F) as a function of F because it is obscured by the cos $2\pi F$ term in the frequency response; however, in the polynomial domain, such a ripple would appear as sketched in Fig. 7. The decision to place this ripple just outside of the passband was arrived at after a numerical search through a set of polynomials that had exactly n_p ripples in the passband and exactly $n_s - 1$ ripples in the stopband so that there is one "invisible" ripple of unconstrained amplitude lying outside the interval (-1, 1). A typical such polynomial is sketched in Fig. 8. The narrowest transition band was found to occur when the "invisible" ripple was brought right up to the passband edge.

Numerical experience with many examples has led the authors to speculate that if the maximally equiripple filter having exactly n_p ripples in the passband and exactly n_s ripples in the stopband is not optimum, then the greatest lower bound of all transition bandwidths in the class can not be achieved by a filter in the class. It appears however that this greatest lower bound can be approached by filters that have one or more invisible ripples just outside the one band edge and that same number of ripples missing from the opposite band.

To partially illustrate a mechanism for this behavior consider Fig. 9 which shows transition bandwidth vs. n_p for fixed n = 10 with $n = n_p + n_s + 1$ so that no ripples are missing. The passband ripple δ_p is 0.01 and the stopband ripple δ_s is 0.0001. We note that as a ripple is exchanged from the stopband to the passband the width of the transition band decreases. Let us consider a specific example with $n_p = 6$ and $n_s = 3$. For this case the width of the transition band is 0.139. If $n_p = 7$ and $n_s = 2$ then the transition bandwidth has been reduced to 0.130 but the number of passband ripples has been increased by one. To remove this ripple the polynomial can be linearly scaled. As the amount of scaling increases the transition width increases. To remove the ripple from the passband, then with the minimum amount of scaling, this ripple should be scaled to the edge of the passband. In certain cases the numbers are such that the decrease in transition bandwidth produced by the removal of a stopband ripple is greater than the increase produced by the subsequent scaling thus resulting in a filter that is better than the maximally equiripple filter having the maximum allowable number of ripples.

REFERENCES

1. O. Herrmann, W. Schuessler, "Design of Nonrecursive Digital Filters with Minimum Phase", Electronics Letters, Vol. 6 No. 11 (May 1970).

2. H. D. Helms, "Digital Filters with Equiripple or Minimax Responses", IEEE Trans. Audio and Electroacoustics, AU-19, No. 1 (March 1971).

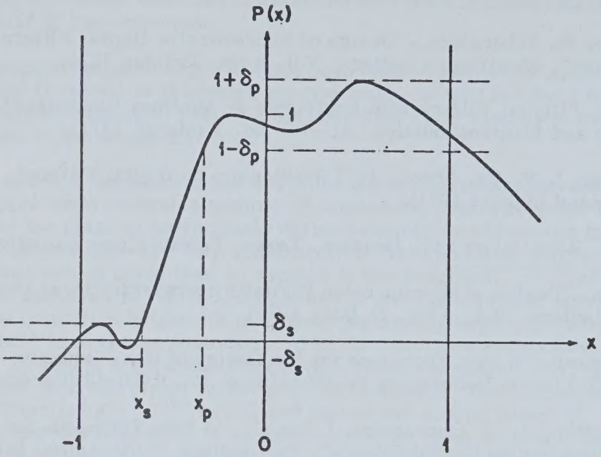
3. L. R. Rabiner, J. V. Hu, "Proc. 1971 Symposium on Digital Filters", Imperial College, England (August 1971).

4. T. W. Parks, Rice University, Houston, Texas, Private Communication.

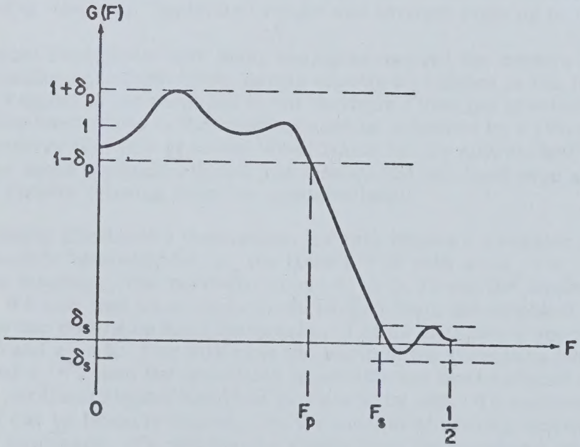
5. O. Herrmann, "Design of Nonrecursive Digital Filters with Linear Phase", Electronics Letters, Vol. 6 No. 11 (May 1970).

6. E. M. Hofstetter, "A New Technique for the Design of Non-Recursive Digital Filters", MIT Lincoln Laboratory Technical Note No. 1970-42 (December 1970).

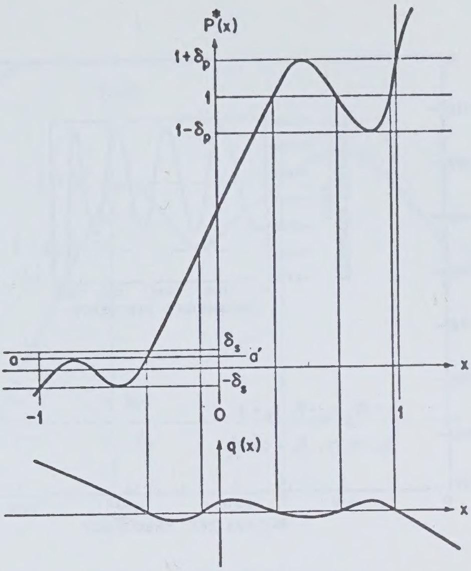
7. E. M. Hofstetter, A. V. Oppenheim, J. Siegel, "A New Technique for the Design of Nonrecursive Digital Filters", Proceedings, Fifth Annual Princeton Conference on Information Sciences and Systems (March 1971).



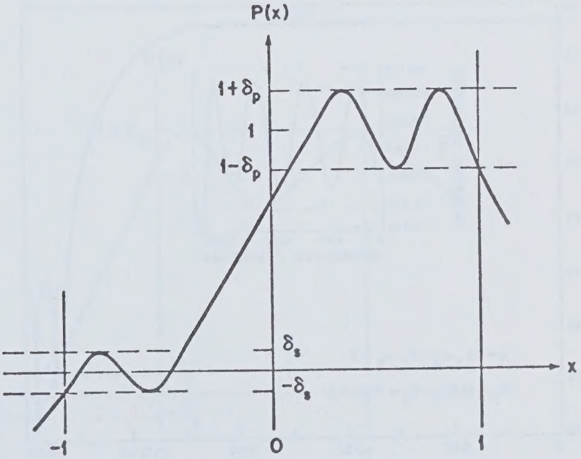
ACCEPTABLE POLYNOMIAL



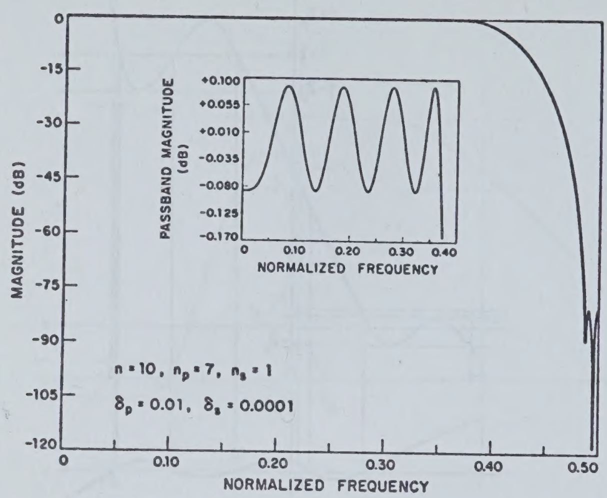
ACCEPTABLE FILTER RESPONSE


P*(x)
1+δp
1
1-δp
δs
o'
o
-δs
-1
0
1
x
q(x)
x

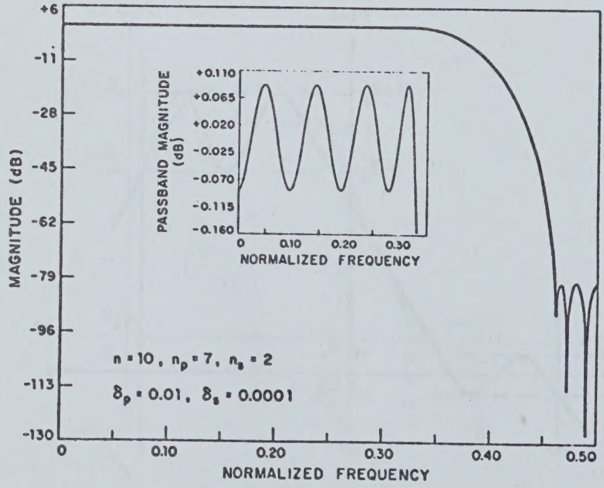
ILLUSTRATING THE CONSTRUCTION OF q(x)


P(x)
1+δp
1
1-δp
δs
-δs
x
-1
0
1

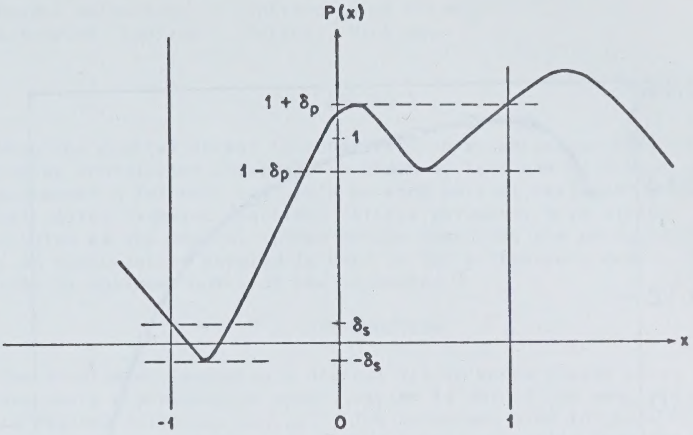
MAXIMALLY EQUIRIPPLE FILTER



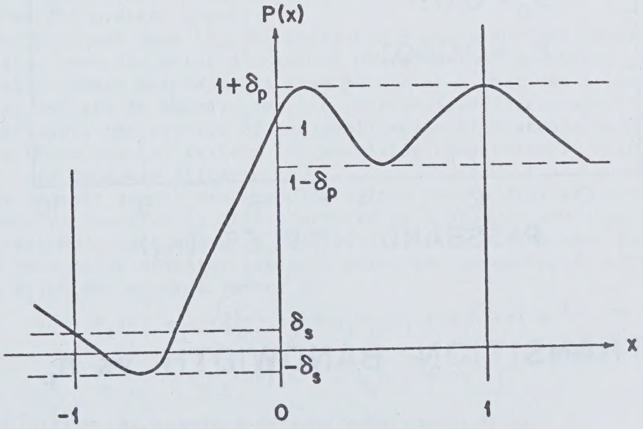
FILTER WITH ONE RIPPLE MISSING



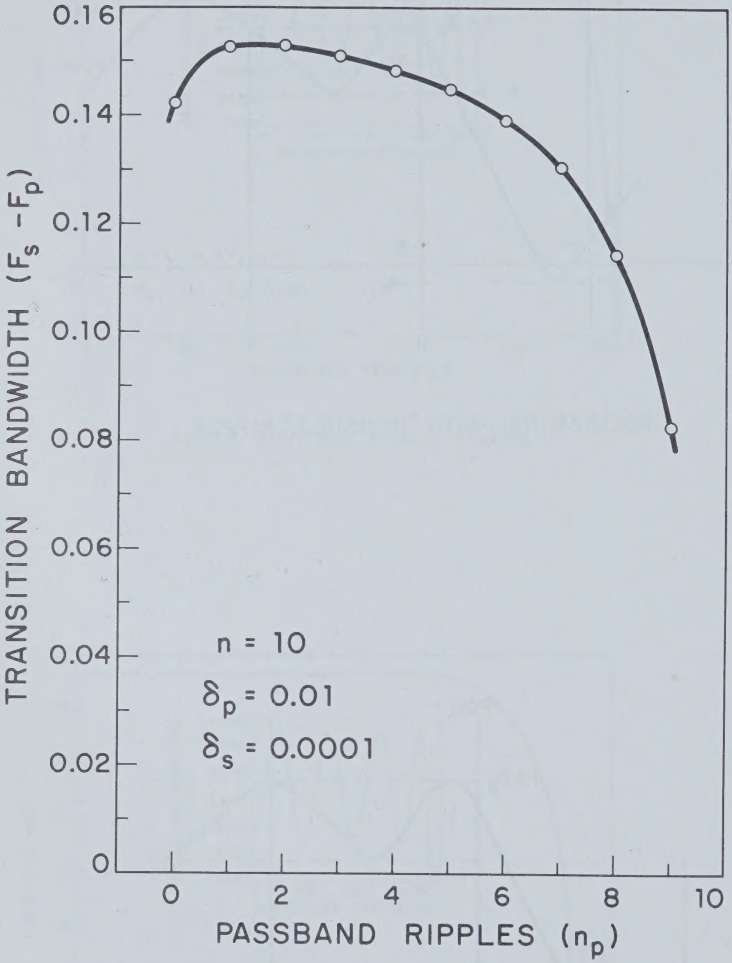
FILTER WITH ALL ALLOWABLE RIPPLES


P(x)
1 + δp
1
1 - δp
δs
-δs
x
-1
0
1

POLYNOMINAL WITH "INVISIBLE" RIPPLE


P(x)
1+δp
1
1-δp
δs
-δs
x
-1
0
1

POLYNOMIAL WITH ENDPOINT RIPPLE



TRANSITION BANDWIDTH vs n_p

DIGITAL FILTERING WITH VARIOUS INTEGRATION ROUTINES

B. J. LEON
Purdue University
Lafayette, Indiana

T-C HUANG
University of Detroit
Detroit, Michigan

ABSTRACT

When the digital filter is considered as a special purpose computer for integrating convolution integrals as shown by Leon and Harrison[6], there is a free parameter δ for each real pole or each pair of conjugant poles in the original filter transfer function. If the parameter δ is varied, the z domain poles of the digital filter remain fixed but the zeros move. In this paper, an optimization program is used to fit a frequency domain filter specification by optimum choice of the parameter δ.

INTRODUCTION

The problem of designing a digital filter whose steady state frequency response meets a preassigned specification is one of the most discussed problems in digital filtering [1],[2]. The technique used in these two books and a large number of papers is to find a transformation from the transfer function of an analog filter that meets the specifications to the pulse transfer function of a digital filter. The standard z-transformation, the matched z-transformation, and the bilinear z-transformation are three in common use.

Steiglitz [3] investigated the basic problem by using optimization. His approach was to consider the second order pulse transfer function with the coefficients of the numerator and denominator polynomials as free parameters. He then applied the Fletcher-Powell [4] method to determine these parameters. When the unconstrained optimization lead to an unstable filter, Steiglitz devised a method for generating a stable filter that was just as good. In fact, he found that applying the optimization procedure to the stable filter sometimes lead to further improvement. This indicated that the originally found unstable filter represented only a local minimum of the error functional. This local minimum problem is also inherent in the method of the present paper.

Harrison and Leon [5],[6] presented a technique for considering a digital filter from the point of view of numerical integration of convolution integrals. Their method has a free parameter δ that can be used in an optimization routine to improve the fit to a desired filter specification. This paper presents the results of the application of Fletcher-Powell optimization to these digital filters for modelling comparatively high order analog lowpass and bandpass filters.

To convert an n^{th} order analog filter into a digital filter the original transfer function is first factored as a product of first and second order rational functions with real coefficients. Then each factor is converted to a pulse transfer function with free parameter δ. For each analog factor $H_i(s)$ the digital factor is

$$K_i(z) = T\left\{\mathcal{Z}\left[H_i(s)\right] + \gamma_1 h_i(0) + \gamma_2 h_i(T)\, z^{-1} + \gamma_3 h_i(2T)\, z^{-2} + \gamma_4 h_i(3T)\, z^{-3}\right\},$$

where $\mathcal{Z}\left[H(s)\right]$ = the standard z-transformation of $H_i(s)$

$h_i(t) = \mathcal{L}^{-1}\left[H_i(s)\right]$ = the inverse-Laplace transform of $H_i(s)$

T = the sampling period

γ_j = functions of the free parameters δ_i as shown in Table 1 for various integration routines. The derivation of this table is given in [5].

TABLE I. Relations between γ_i and δ.

Integration routine		γ_1	γ_2	γ_3	γ_4
Rectangular $0<\delta<1$		δ	0	0	0
Quadratic $1<\delta<2$	QIa	$\frac{1}{2}\delta^2-2\delta-11/12$	$-\frac{1}{2}\delta^2+\delta-5/12$	0	0
	QIb	$\frac{1}{2}(\delta^2-3\delta+5/6)$	$\frac{1}{2}(-\delta^2+3\delta-\frac{11}{6})$	0	0
	QII	$-7/12$	$1/12$	0	0
	QIII	$(2\delta^3-9\delta^2+12\delta-12)/12$	$(-4\delta^3+18\delta^2-24\delta+11)/12$	$\frac{(2\delta^3-9\delta^2+12\delta-5)}{12}$	0
Cubic $1<\delta<3$	CuI	$(4\delta^3-30\delta^2+72\delta-31)/24$	$(8\delta^3-48\delta^2+72\delta-28)/24$	$\frac{4\delta^3-18\delta^2+24\delta-9}{24}$	0
	CuII	$-5/8$	$1/6$	$-1/24$	0
	CuIII	$(-\delta^4+8\delta^3-22\delta^2+24\delta-24)/24$	$(3\delta^4-24\delta^3+66\delta^2-72\delta+31)/24$	$(-3\delta^4+24\delta^3-66\delta^2+72\delta-28)/24$	$\frac{(\delta-1)^2(\delta-3)^2}{24}$

In the examples below we consider various Butterworth filters. In the odd order cases there is one real pole and $\frac{n-1}{2}$ pole pairs. In the even order cases there are $\frac{n}{2}$ pairs and no real poles. Thus, for an n^{th} order lowpass filter there are $\frac{n+1}{2}$ or $\frac{n}{2}$ parameters δ_i for the optimization routine. For an n^{th} order bandpass filter there are n pole pairs and n parameters δ_i. Since Steiglitz [3] had four free parameters per second order section, the present method is computationally simpler. As seen below, it does give very good results on the examples chosen.

CRITERION FOR OPTIMIZATION

The optimization is applied to obtain the digital filter whose steady state response approximates the frequency characteristics of an analog filter. The criterion chosen is the minimum of the sum of squares of the error at a set of discrete frequencies. Let $H(s)$ be the transfer function of an analog filter, and ω_i, $i=1,2,\ldots,n_f$, be the chosen set of discrete frequencies. Define

$$H_i = 10 \log \frac{|H(j\omega_i)|^2}{H^2(0)} \tag{2}$$

The criterion adopted for the optimum of the pulse-transfer function of the digital filter is to minimize the total error defined as

$$E(\underline{\delta}) = \sum_{i=1}^{n_f} \left[AMP(j\omega_i) - H_i \right]^2 \tag{3}$$

where

$$AMP(j\omega_i) = 10 \log \frac{|K(e^{j\omega_i T})|}{K^2(1)} \tag{4}$$

$K(z)$ is the overall pulse transfer function,

$\underline{\delta}$ = the vector of free parameters δ_i.

The problem here is to find $\underline{\delta}$ so that $E(\underline{\delta})$ is minimized. The $\underline{\delta}$ is, then, a set of optimal values in the least square error sense.

OPTIMIZATION PROCEDURE

A computer program implementing the Fletcher and Powell technique by D. A. Calahan [7] is adapted with minor modifications for our problem previously stated. For the program, the computation of the gradient of the error function $E(\underline{\delta})$ is done by numerical methods which are less complicated than analytical minimization. Analytical minimization is used to help determine the initial values of parameters δ_m.

Differentiating $E(\underline{\delta})$ with respect to parameter δ_m, we have

$$\frac{\partial E(\underline{\delta})}{\partial \delta_m} = 2 \sum_{i=1}^{\frac{1}{2}(n+1) \text{ or } \frac{n}{2}} (AMP(\omega_i) - H_i) \frac{\partial AMP(\omega_i)}{\partial \delta_m} \qquad (5)$$

$$j=1,2,\ldots,\tfrac{n}{2} \text{ (or } \tfrac{1}{2}(n+1))$$

Differentiating Eq. (4) gives

$$\frac{\partial AMP(\omega_i)}{\partial \delta_m} = \frac{20}{|K(e^{j\omega_i T})|^2} \operatorname{Re}\left\{ K(e^{j\omega_i T}) \left[\frac{\partial \gamma_{m1}}{\partial \delta_m} h_m(0) + \frac{\partial \gamma_{m2}}{\partial \delta_m} h_m(T)\, e^{-j\omega_i T} + \frac{\partial \gamma_{m3}}{\partial \delta_m} h_m(2T)\, e^{-j2\omega_i T} + \frac{\partial \gamma_{m4}}{\partial \delta_m} h_m(3T)\, e^{-j3\omega_i T} \right] \right\} \qquad (6)$$

where $\frac{\partial \gamma_{mi}}{\partial \delta_m}$ is given in Table II.

TABLE II. The Derivatives of γ_i

Design Types		$\frac{\partial \gamma_1}{\partial \delta}$	$\frac{\partial \gamma_2}{\partial \delta}$	$\frac{\partial \gamma_3}{\partial \delta}$	$\frac{\partial \gamma_4}{\partial \delta}$
Rectangular 0 δ I		-1	0	0	0
Quadratic $1<\delta<2$	QIa	$\delta - 2$	$1 - \delta$	0	0
	QIb	$\delta - 3/2$	$-\delta + 3/2$	0	0
	QII	0	0	0	0
	QIII	$\frac{1}{2}(\delta-1)(\delta-2)$	$-(\delta-1)(\delta-2)$	$\frac{1}{2}(\delta-1)(\delta-2)$	0
Cubic $1<\delta<3$	CuI	$-\frac{1}{2}(\delta-3)(\delta-2)$	$(\delta-3)(\delta-1)$	$-\frac{1}{2}(\delta-2)(\delta-1)$	0
	CuII	0	0	0	0
	CuIII	$-(\delta-1)(\delta-2)(\delta-3)/6$	$(\delta-1)(\delta-2)(\delta-3)/2$	$-\frac{1}{2}(\delta-1)(\delta-2)(\delta-3)$	$-(\delta-1)(\delta-2)(\delta-3)/6$

To choose the starting values of $\underline{\delta}$ for the search for the optimum of $E(\underline{\delta})$ we examine Eq. (5) for simple cases.

Now let m=1; Eq. (5) becomes

$$\frac{\partial E}{\partial \delta} = 2 \sum_{i=1}^{n_f} (AMP(\omega_i) - H_i) \frac{\partial AMP(\omega_i)}{\partial \delta} \qquad (7)$$

where

$$\frac{\partial AMP(\omega_i)}{\partial \delta} = -20 \frac{\mathrm{Re}\left\{K(e^{-j\omega_i T})\right\}}{\left|K(e^{j\omega_i T})\right|^2} h(0) \qquad \text{for rectangular integration}$$

$$= \frac{20}{\left|K(e^{j\omega_i T})\right|^2} \mathrm{Re}\left\{K(e^{-j\omega_i T}) \left[(\delta-2)h(0)+(1-\delta)\, h(T)\, e^{-j\omega_i T}\right]\right\} \qquad \text{for quadratic integration Ia}$$

$$= \frac{20}{\left|K(e^{j\omega_i T})\right|^2} \mathrm{Re}\left\{K(e^{-j\omega_i T}) \left[(\delta-3/2)\, h(0)+(\tfrac{3}{2} - \delta)\, e^{-j\omega_i T}\right]\right\} \qquad \text{for quadratic integration Ib}$$

$$= \frac{20}{\left|K(e^{j\omega_i T})\right|^2} \mathrm{Re}\left\{\tfrac{1}{2}(\delta-1)(\delta-2)\, K(e^{-j\omega_i T}) \left[h(0) - 2h(T)\, e^{-j\omega_i T} + h(2T)\, e^{-j2\omega_i T}\right]\right\} \qquad \text{for quadratic integration III}$$

$$= \frac{20}{\left|K(e^{j\omega_i T})\right|^2} \mathrm{Re}\left\{K(e^{-j\omega_i T}) \left[-\tfrac{1}{2}(\delta-3)(\delta-2)\, h(0) + (\delta-3)(\delta-1)\, h(T)\, e^{-j\omega_i T}\right]\right\} \qquad \text{for cubic integration I}$$

$$= \frac{20}{\left|K(e^{j\omega_i T})\right|^2} \mathrm{Re}\left\{\frac{(\delta-1)(\delta-2)(\delta-3)}{6} K(e^{-j\omega_i T}) \left[-h(0) + 3h(T)\, e^{-j\omega_i T} - 3h(2T)\, e^{-j2\omega_i T} - h(3T)\, e^{-j3\omega_i T}\right]\right\} \qquad \text{for cubic integration III.}$$

From the above expressions we observe the following:

1) For the quadratic design QIa, if h(0)=0, there is an extremum at $\delta=1$, independent of the set of discrete frequencies chosen.
2) For the quadratic design QIb, there is one extremum at 1.5.
3) For the quadratic design QIII, there are three extrema at $\delta=1,2$, and 3.
4) For the cubic design CuIII, there are extrema at $\delta=1,2$, and 3,too.

Though these remarks do not solve the problem of minimizing $E(\underline{\delta})$, they help us select a preferable set of starting values of $\underline{\delta}$ in searching for the optimum of $E(\underline{\delta})$.

For a first order system, the loci of parameter δ at which $E(\delta)$ is minimum are studied in detail in order to investigate the properties of these extrema. The transfer function for this study is a first order Butterworth filter, $\frac{10}{s+6.2831853}$, and the design types QIb and CuIII are taken to examine the extrema mentioned in remarks 2) and 3). The loci of the minima of $E(\underline{\delta})$ shown in Fig. 1 indicate that these extrema may be minima for certain values of T, and maxima at other values of T. So, these extrema are useful only in picking the starting values of the parameters δ for the Fletcher-Powell algorithm.

DESIGN EXAMPLES

Lowpass Filter

For illustration, a digital filter having a frequency characteristic of a 9th order Butterworth filter is designed. Suppose that the sampling period T=0.1 sec. and the cutoff frequency of the lowpass filter is 1Hz. The first step in the design procedure is to find the transfer function of the analog filter corresponding to the Butterworth filter of the 9th order from filter tables [8].

$$H(s) = \frac{10}{s+b_{11}} \prod_{i=2}^{5} \frac{10}{s^2+b_{i2}s+b_{i1}} = \prod_{i=1}^{5} H_i(s)$$

where

$$H_i(s) = \frac{10}{s^2+b_{i2}s+b_{i1}} \qquad \text{for } i \neq 1$$

$$= \frac{10}{s + b_{11}} \qquad \text{for } i=1$$

and

$b_{11} = 6.2831853$

$b_{21} = 39.4784175$, $b_{22} = 2.1821276$

$b_{31} = 39.4784175$, $b_{32} = 6.2831853$

$b_{41} = 39.4784175$, $b_{42} = 9.6263978$

$b_{51} = 39.4784175$, $b_{52} = 11.8285256$

Then, applying the transformation technique to each factor $H_i(s)$, we have a general expression for the pulse-transfer function of the digital filter.

$$K(z) = \prod_{i=1}^{5} \left[\frac{c_{i1} + c_{i2}z^{-1}}{1 + d_{i2}z^{-1} + d_{i3}z^{-2}} + \gamma_{i1}\, h_i(0) + \gamma_{i2}\, h_i(T)\, z^{-1} + \gamma_{i3}\, h_i(2T)\, z^{-2} + \gamma_{i4}\, h_i(3T)\, z^{-3} \right]$$

where

$h_i(t)$ = the inverse Laplace transform of $H_i(s)$,

γ_{ik} are determined by the design type chosen as in Table 1,

and

$c_{11} = 10$, $c_{12} = 0$, $d_{12} = -0.5334881$, $d_{13} = 0$,

$c_{21} = 0$, $c_{22} = 0.8405032$, $d_{22} = -1.4607813$, $d_{23} = 0.8039544$,

$c_{31} = 0$, $c_{32} = 0.6948886$, $d_{32} = -1.2498255$, $d_{33} = 0.5334881$,

$c_{41} = 0$, $c_{42} = 0.6013037$, $d_{42} = -1.1364972$, $d_{43} = 0.3818835$,

$c_{51} = 0$, $c_{52} = 0.5498361$, $d_{52} = -1.0826919$, $d_{53} = 0.3070169$.

There are five cascades. So, there are at least five free parameters δ_i. The criterion for the optimization is the minimum of total square error over the frequencies ω = 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, and 1.8 Hz. The optimal set of parameters δ_i are searched for various design types as shown in Table III. Note that for the rectangular integration, there is only one effective parameter δ_i in the function K(z), since $h_i(0) = 0$ for i = 2, 3, 4, and 5. The tables show that quadratic integration and cubic integration do not always appear to be better than rectangular integration. Among these design types, rectangular design for the first cascade along with cubic design III for the rest appears to be the "best." Some optimal sets of δ_i are found to be out of the range originally allowed by Harrison and Leon. However, these values still lead to realizable filters. The analog filter response along with the standard z-transformation filter and the best design from Table III are plotted in Fig. 2.

TABLE III. The Optimal Values Of Parameters δ For The 9th Order Butterworth Lowpass Filter

1st cascade of the transfer function $\dfrac{10}{s + 6.2831853}$

2nd cascade of the transfer function $\dfrac{10}{s^2 + 2.1821276s + 39.4784175}$

3rd cascade of the transfer function $\dfrac{10}{s^2 + 6.2831853s + 39.4784175}$

4th cascade of the transfer function $\dfrac{10}{s^2 + 9.6263978s + 39.4784175}$

5th cascade of the transfer function $\dfrac{10}{s^2 + 11.8285256s + 39.4784175}$

Cascade-Design Type*	Starting Value of δ	Optimal Value of δ	Total Square Error
1 Rect.	0.5	0.65210839	17.5296881
1QIa	1.6	1.08583909	21.8829848
1QIb	1.4	0.64305313	17.6254072
	1.8	2.35699674	17.6254071
1QIII	1.6	1.00001211	22.3177386
ICuI	1.6	1.03379713	26.3931876
	2.2	2.65231379	31.0386988
1CuIII	1.6	0.99999997	26.4364053
	2.2	2.99993229	26.4364054
1QIa	1.1	1.08760718	13.8324477
2QIa	1.9	1.00475033	
3QIa	1.9	1.30057405	
4QIa	1.1	1.00283944	
5QIa	1.1	1.69437009	
IQIb	1.4	0.65828870	4.8556712
2QIb	1.4	1.50097650	
3QIb	1.4	0.88338382	
4QIb	1.4	1.48429857	
5QIb	1.4	0.62372283	

Cascade-Design Type*	Starting Value of δ	Optimal Value of δ	Total Square Error
1QIII	1.6	1.01353937	1.2346744
2QIII	1.6	1.98279733	
3QIII	1.6	1.97321358	
4QIII	1.6	1.98110266	
5QIII	1.6	2.01104125	
1CuI	1.8	1.02827723	9.5070473
2CuI	1.8	1.14936737	
3CuI	1.8	1.14206242	
4CuI	1.8	1.07011734	
5CuI	1.8	1.39166809	
1CuIII	2.2	2.96483723	0.0619482
2CuIII	2.2	2.62575388	
3CuIII	2.2	1.92877164	
4CuIII	2.2	2.02671152	
5CuIII	2.2	1.89902567	
1Rect.	0.6	0.66224503	0.0358706
2CuIII	1.4	1.44773175	
3CuIII	1.4	1.50160214	
4CuIII	1.4	1.49011713	
5CuIII	1.4	1.50410126	
1Rect.	0.6	0.64992366	0.3979443
2QIII	1.8	1.99640086	
3QIII	1.8	1.99416675	
4QIII	1.8	1.99579213	
5QIII	1.8	2.00798274	

*The cascades omitted in each combination were taken by conventional z-transformation with no optimization.

A more dramatic improvement in the frequency characteristics after optimization occurs when the ratio of the cutoff frequency of the lowpass filter to the sampling frequency is increased to 0.25 sec., we apply the quadratic design Ia to the first three cascades and the quadratic design Ib to the last two, and find the optimal set of δ_i. The result is shown in Fig. 3.

Bandpass Filters

To design bandpass filters, we proceed as follows:

1) Design a prototype lowpass analog filter.
2) Transform the lowpass filter into the desired bandpass filter.
3) Use the approximating integral convolution technique to transform the analog filter into the digital filter.

Steps (1) and (3) are done as for lowpass filters. Step (2) is accomplished either by shifting the pole-zero configuration of the lowpass filters or by the conventional technique through replacing s in the transfer function of the analog filter by $(s^2 + \omega_0^2)/Bs$ where ω_0 is the center frequency of the pass band and B, the bandwidth. One of the features of the conventional transformation is that, for each finite pole (or zero), it introduces a pair of poles (or zeros) with one zero (or pole) at $s=0$ in the s-plane. The shifting technique is simply to shift each pole (or zero) along the $j\omega$-axis by ω_0 up and down to form a pair of new poles (or zeros). After the resulted filter is transformed into digital form through step (3), the pole-zero configuration in the z-domain is the same as that of the lowpass filter formed from the prototype lowpass analog filter, except for a rotating angle $e^{j\omega_0 T}$. The cutoff of the low frequency portion is not as sharp as that derived from the conventional transformation. To sharpen it, some proper zeros must be introduced.

A 3rd order Butterworth bandpass filter is used to show the effectiveness of the optimization. The center frequency is set at 2.5 Hz, the sampling period $T=0.1$ sec., and the frequencies ω_i in the minimization criterion are 1.5, 2.0, 2.2, 2.4, 2.5, 2.6, 2.8, 3.0, 3.5, and 4.0 Hz. The optimal set of $\underline{\delta}$ for digital filters derived from conventional transformation and from shifting the pole-zero configuration with zeros added at $s=0$ are shown in Table IV. Moreover, among the design types, the quadratic integration QIII and cubic integration CuIII are less favorable, while the cubic design CuI is the best. Figures 4 and 5 show the cubic design CuI.

TABLE IV. Optimal Set Of δ For Bandpass Filters

	a. Conventional Transformation	b. Shifting Pole-Zero Configuration with Zeros Added
1st Cascade	$\frac{10s}{s^2 + 6.2831853s + 246.7401093}$	$\frac{10s}{s^2 + 6.2831853s + 256.6097318}$
2nd Cascade	$\frac{10s}{s^2 + 2.6028916s + 174.5072023}$	$\frac{10s}{s^2 + 3.1415927s + 171.1364330}$
3rd Cascade	$\frac{10s}{s^2 + 3.6802937s + 348.8720283}$	$\frac{10s}{s^2 + 3.1415927s + 342.0829946}$

Cascade-Design Type	a			b		
	Starting Value	Optimal Value	Total Square Errors	Starting Value	Optimal Value	Total Square Errors
1 Rect.	0.5	0.33645944	22.2055128	0.6	0.35665863	20.3183341
2 Rect.	0.5	-0.70099714		0.6	-0.85052045	
3 Rect.	0.5	0.48931260		0.6	0.48299740	
1QIa	1.3	2.81077188	51.8949376	1.4	2.81815364	41.9632666
2QIa	1.4	1.85712179		1.4	1.82214142	
3QIa	1.4	0.71042247		1.4	0.69607630	
1QIb	1.6	2.31549536	23.2658782	1.6	2.29289427	20.3284679
2QIb	1.6	-0.08825493		1.6	-0.16761033	
3QIb	1.6	2.22021641		1.6	2.22414160	
1QIII	1.6	0.96852480	313.6249486	1.6	0.98454889	318.3075230
2QIII	1.6	1.01177538		1.6	1.00139273	
3QIII	1.6	0.97542588		1.6	0.97808874	
1CuI	1.8	2.84871128	3.0635158	1.8	2.88438536	2.8860816
2CuI	1.8	1.65167687		1.8	1.69384281	
3CuI	1.8	2.84518461		1.8	2.83828488	
1CuIII	1.8	0.95996751	396.0236654	1.6	1.0476688	404.7462506
2CuIII	1.8	0.98797521		1.6	1.00038336	
3CuIII	1.8	0.99881112		1.6	0.98875374	
1CuIII	2.2	3.04061660	396.0359500	2.2	3.00934507	404.7146890
2CuIII	2.2	3.01251098		2.2	0.99468514	
3CuIII	2.2	3.00121336		2.2	0.99902415	
1 Rect.	0.5	0.17084108	63.2279269	0.5	0.19842811	63.0320283
1QIa	1.5	2.85305059	42.8084906	1.5	2.84607747	32.1088159
1QIb	1.1	0.46998430	62.2694952	1.1	0.50045451	61.9229706
1QIb	1.8	2.53001819	62.2694952	1.8	2.49954517	61.9229706
1QIII	1.8	1.00000403	95.2312076	1.8	1.00000796	92.3489896
1CuI	1.8	2.86034742	58.3585861	1.8	2.90242869	47.6863864
1CuIII	2.0	2.54000000	106.2478771	2.0	2.52459738	102.8468770

CONCLUSION

After an optimization program is implemented, the optimal set of free parameters δ_i for a chosen design type of a digital filter can be easily found, though a few attempts may be needed because of the existence of several local minima. When the design type is not specified, the choice of the design type depends, besides the approximation criterion, on the hardware realization. The rectangular design has a simple hardware configuration, while the cubic design requires more shift registers, adders and multipliers. So, economically speaking, it is slightly more desirable to use the rectangular integration, if the specifications of the design are not too strict. But, as pointed out in [6] for high order filters the added hardware required for quadratic and cubic integration is not very significant.

REFERENCES

[1] J. F. Kaiser, "Digital Filters," in System Analysis by Digital Computer, F. F. Kuo and J. F. Kaiser, Eds., New York, Wiley, 1966.

[2] B. Gold and C. M. Raker, Digital Processing of Signals, New York, McGraw-Hill, 1969.

[3] K. Steiglitz, "Computer-Aided Design of Recursive Digital Filters," IEEE Trans. AV-18, pp. 123-29, June 1970.

[4] R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization," Computer J, vol. 6, pp. 163-8, 1963.

[5] S. R. Harrison and B. J. Leon, "Digital Filters," Technical Report TREE 69-35, School of EE, Purdue Univ., September 1969.

[6] S. R. Harrison and B. J. Leon, "Digital Filters for Approximating Continuous Convolutions," IEEE Trans. CT-18, November 1971.

[7] D. A. Calahan, Computer-Aided Network Design, New York, McGraw-Hill, 1968.

[8] L. Weinberg, Network Analysis and Synthesis, New York, McGraw-Hill, 1962.

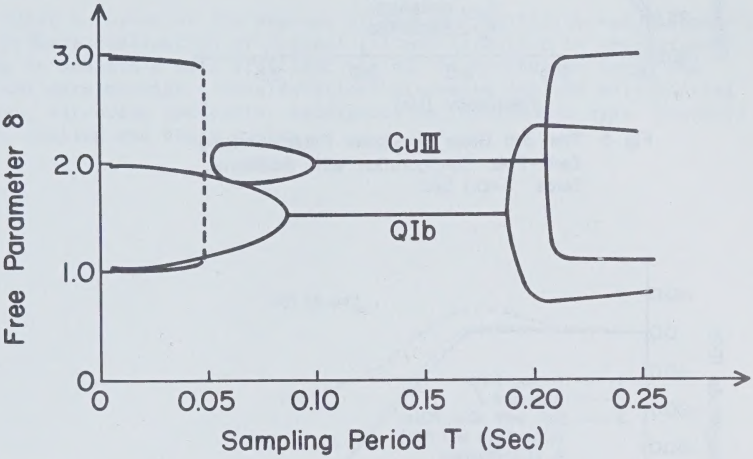


Fig. 1 Minimal Loci of the Free Parameter δ for the First Order System $H(s) = \frac{10}{s+6.2831853}$

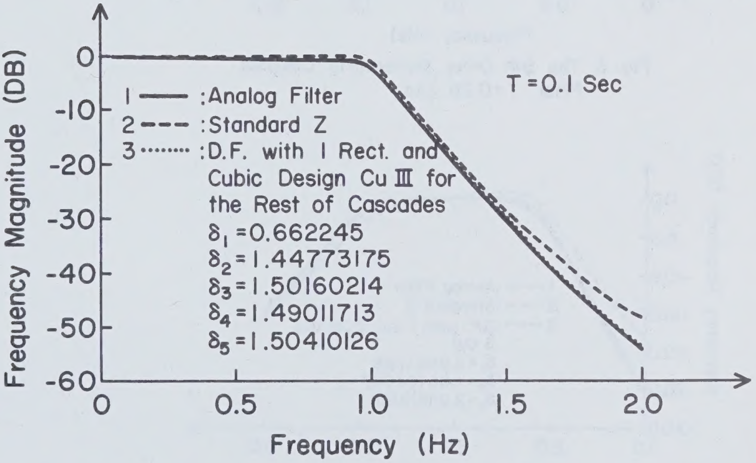


Fig. 2 The 9th Order Butterworth Lowpass Filter T = 0.1 Sec.

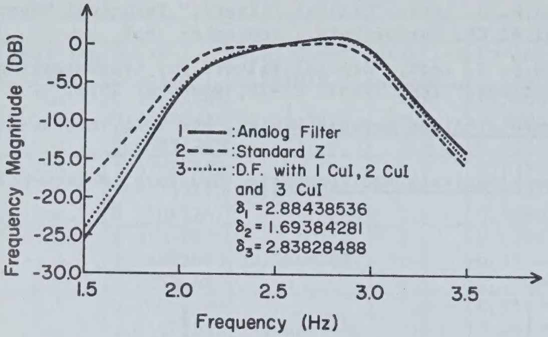


Fig. 5 The 3rd Order Bandpass Filter - Shifting Zero-Pole Configuration with Additional Zeros. T = 0.1 Sec.

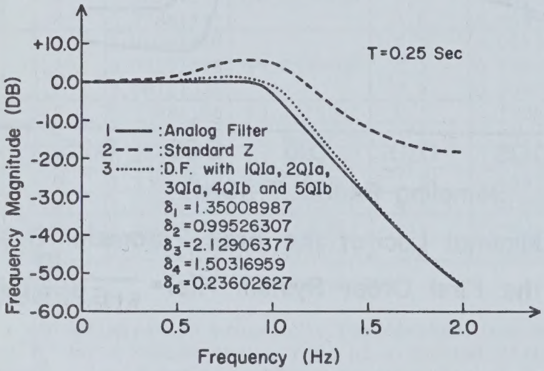


Fig. 3 The 9th Order Butterworth Lowpass Filter T = 0.25 Sec.

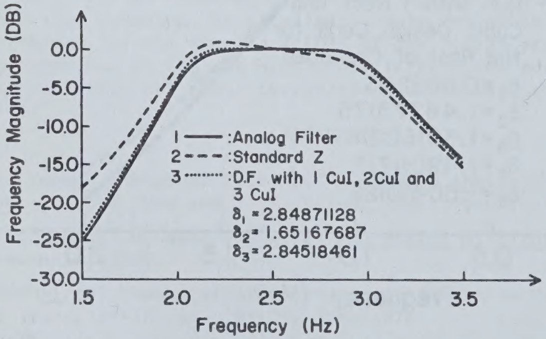


Fig. 4 The 3rd Order Bandpass Filter - Conventional Transformation. T = 0.1 Sec.

SOME DESIGN CONSIDERATIONS IN THE IMPLEMENTATION OF DIGITAL FILTERS

J. F. Kaiser
Bell Telephone Laboratories
Murray Hill, N. J.

Abstract

After a review of the degrees of design flexibility associated with the hardware realization of digital filters attention is concentrated on trying to achieve a most efficient use of the arithmetic logic and internal data storage. Considerations discussed include multiplexing schemes, bit-order tradeoffs, configuration, arithmetic type, roundoff noise, scaling and limit cycles.

ON CANONIC REALIZATIONS OF DIGITAL FILTERS

S. K. Mitra and R. Sherwood
University of California
Davis, Calif. 95616

ABSTRACT

Canonic realizations of a digital transfer function using the continued fraction expansion techniques are derived. Realizability conditions of each realization are provided along with illustrative examples.

INTRODUCTION

With the rapid advances in digital integrated circuit technology, there has been an increasing interest in digital signal processing in recent years as this approach provides an alternate solution to micro-miniaturized inductorless filter realization. Practical constraints associated with the physical implementation of a digital filter usually cause the actual performance of the filter to deviate from the desired performance. The two major sources of error are the round-off of the coefficients of the digital transfer function and the round-off of the results of the associated multiplication processes. For specific cases, it has been shown that different realizations have different error properties [1-3]. It is therefore desirable to synthesize all possible realizations of a given function, so that an error analysis of each model can be made. In some cases this will lead to implementations with lower errors.

The purpose of this paper is to provide several new canonic realizations of a digital transfer function. The realizations are based on various continued fraction expansions of the transfer function. Realizability conditions of each form are also included along with illustrative examples.

CANONIC REALIZATIONS

The digital transfer function can be expressed as a real rational function in the complex variable z:

$$G(z) = \frac{a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \dots + a_1 z + a_o}{z^n + b_{n-1}z^{n-1} + \dots + b_1 z + b_o} \qquad (1)$$

where we have assumed without loss of any generality that the degree of the numerator polynomial of G(z) is one less that that of the denominator polynomial. If the polynomials are of equal order, an appropriate multiplier may be connected directly from the input to the output summer, and the remaining function will be of the form of Equation (1). Four different canonic realizations are proposed here, each leading to a realization resembling a ladder. The various realizations are discussed next.

Type IA Ladder Form

The first form is realized by performing a continued fraction expansion of G(z) in the form of Equation (2):

$$G(z) = \cfrac{1}{B_n z + \cfrac{1}{A_n + \cfrac{1}{B_{n-1} z + \cfrac{1}{\ddots + \cfrac{1}{B_1 z + \cfrac{1}{A_1}}}}}} \tag{2}$$

where $\{A_i\}$, $\{B_i\}$ are real constants.

In order to implement a function of this form, we need building blocks which realize the two functions:

$$G_1(z) = \frac{1}{Bz + T(z)} \tag{3}$$

$$G_2(z) = \frac{1}{A + T(z)} \tag{4}$$

where A and B are real. Figure 1 shows a realization of $G_1(z)$ and Figure 2 shows two possible realizations of $G_2(z)$. The realization of G(z) as given by Equation (2) is then obtained as follows. We first write G(z) as:

$$G(z) = \frac{1}{B_n z + T_1(z)} \tag{5}$$

where

$$T_1(z) = \cfrac{1}{A_n + \cfrac{1}{B_{n-1} z + \cfrac{1}{A_{n-1} + \cfrac{1}{\ddots + \cfrac{1}{B_1 z + \cfrac{1}{A_1}}}}}} \tag{6}$$

Note that expression (5) is similar in form to that of $G_1(z)$ and can be realized the way indicated in Figure 1. Next $T_1(z)$ is expressed in the form of $G_2(z)$ and realized accordingly. This process is repeated leading to the complete realization as sketched in Figure 3.

Type IB Ladder Form

Another ladder can be developed by expanding G(z) in the form of Equation (7):

$$G(z) = \cfrac{1}{A_1 + \cfrac{1}{B_1 z^{-1} + \cfrac{1}{A_2 + \cfrac{1}{\ddots + \cfrac{1}{B_n z^{-1}}}}}} \tag{7}$$

It should be noted that the real constants $\{A_i\}$, $\{B_i\}$ in the above equation are not the same as those in Equation (2).

In the implementation of Equation (7) we need building blocks realizing functions of the form:

$$G_2(z) = \frac{1}{A + T(z)} \tag{8}$$

$$G_3(z) = \frac{1}{Bz^{-1} + T(z)} \tag{9}$$

Realization of $G_2(z)$ has already been shown (see Figure 2). $G_3(z)$ can be implemented as shown in Figure 4.

Using a step-by-step procedure, the complete realization of Equation (7) is obtained as sketched in Figure 5.

Type IIA Ladder Form

An alternate realization results from the continued fraction expansion of G(z) as:

$$G(z) = \cfrac{1}{B_1 z + A_1 + \cfrac{1}{B_2 z + A_2 + \cfrac{1}{\ddots + \cfrac{1}{B_n z + A_n}}}} \tag{10}$$

The basic building block in the implementation of Equation (10) shown in Figure 6 realizes the transfer function.

$$G_4(z) = \frac{1}{Bz + A + T(z)} \tag{11}$$

The final realization of Equation (10) is indicated in Figure 7, for n odd. If n is even, the same procedure is used, but the last stage in Figure 7 has a mirror image orientation.

Type IIB Ladder Form

The fourth ladder form is generated by expanding $G(z)$ in the way indicated in Equation (12):

$$G(z) = A_o + \cfrac{1}{B_1 z^{-1} + A_1 + \cfrac{1}{B_2 z^{-1} + A_2 + \cfrac{1}{\ddots + \cfrac{1}{B_n z^{-1} + A_n}}}} \tag{12}$$

The basic building block in this case is the one shown in Figure 8 which implements:

$$G_5(z) = \frac{1}{Bz^{-1} + A + T(z)} \tag{13}$$

With the aid of this building block, the alternate realization of $G(z)$ for n odd as sketched in Figure 9 is readily obtained.

REALIZABILITY CONDITIONS

The synthesis techniques outlined in the previous section may be applied only when the corresponding continued fraction expansion exists. In order to determine the existence of the expansion, the following procedure may be followed.

Type A Realization

From the specified transfer function as given by Equation (1), we form the $2n \times 2n$ matrix $\underline{M}$ as given below:

$$\underline{M} = \begin{bmatrix} a_{n-1} & a_{n-2} & a_{n-3} & \cdots & a_2 & a_1 & a_o & 0 & 0 & \cdots \\ 1 & b_{n-1} & b_{n-2} & \cdots & b_3 & b_2 & b_1 & b_o & 0 & \cdots \\ 0 & a_{n-1} & a_{n-2} & \cdots & a_3 & a_2 & a_1 & a_o & 0 & \cdots \\ 0 & 1 & b_{n-1} & \cdots & b_4 & b_3 & b_2 & b_1 & b_o & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{14}$$

Let us denote the $i^{\underline{th}}$ principal minor of $\underline{M}$ by D_i.

For the existence of a Type IA realization, we must have $D_i \neq 0$ for $i = 1,2,3,\ldots, r$ where $r = 2n-1$ if $b_0 = 0$ otherwise $r = 2n$. For the existence of a Type IIA realization, it is only necessary to insure that the odd numbered principal minors are non-zero, i.e. $D_i = 0$ for for $i = 1,3,5, \ldots , 2n-3, 2n-1$. These conditions are both necessary and sufficient.

Type B Realization

To check the existence of the Type B realization we rewrite $G(z)$ as:

$$G(z) = \frac{a_j z^j + \cdot \cdot \cdot \; a_{n-1} z^{n-1}}{b_k z^k + \cdot \cdot \cdot \; b_{n-1} z^{n-1} + z^n} \tag{15}$$

where $a_i = 0$ for $i < j$ and $b_i = 0$ for $i < k$, i.e. a_j is the first non-zero coefficient in the numerator polynomial and b_k is the first non-zero coefficient in the denominator polynomial of $G(z)$. Next, we express $G(z)$ as:

$$G(z) = c_m z^{-m} + \cdot \cdot \cdot + c_1 z^{-1} + c_o$$
$$+ \frac{d_{k+1} z^{k+1} + d_{k+2} z^{k+2} + \cdot \cdot \cdot + d_{n-1} z^{n-1}}{b_k z^k + b_{k+1} z^{k+1} + \cdot \cdot \cdot + z^n} \tag{16}$$

It should be noted that if $j < k$, then all the c_i in Equation (16) are zero.

There are three possible cases to examine.

Case 1: $j + 1 < k$ - no Type B realization exists.
Case 2: $j + 1 = k$ - $d_i = a_i$ for $i = 0,1, \cdot\cdot , n-1$.
Case 3: $j + 1 < k$ - d_i determined by actual expression of $G(z)$ in the form of Equation (16). If $d_{k+1} = 0$, then also Type B realization does not exist.

For Cases 2 and 3, we then form the $2(n-k) \times 2(n-k)$ $\underline{M}$ matrix:

$$\underline{M} = \begin{bmatrix} d_{k+1} & d_{k+2} & d_{k+3} & \cdots & d_{n-2} & d_{n-1} & 0 & 0 & \cdots \\ b_k & b_{k+1} & b_{k+2} & \cdots & b_{n-3} & b_{n-2} & b_{n-1} & 1 & \cdots \\ 0 & d_{k+1} & d_{k+2} & \cdots & d_{n-3} & d_{n-2} & d_{n-1} & 0 & \cdots \\ 0 & b_k & b_{k+1} & \cdots & b_{n-4} & b_{n-3} & b_{n-2} & b_{n-1} & \cdots \\ 0 & 0 & d_{k+1} & \cdots & d_{n-4} & d_{n-3} & d_{n-2} & d_{n-1} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \end{bmatrix} \tag{17}$$

Let us denote again the $i^{\underline{th}}$ principal minor of $\underline{M}$ by D_i.

For Type I realization, all D_i must be non-zero for $i = 1,2,\ldots 2(n-k)$. For Type II realization, only the odd numbered principal minors must be non-zero. The conditions for the Type IB realization are both necessary and sufficient. For Type IIB realization, we must also insure that the constants A_i in the continued-fraction expansion of $G(z)$ as given by Equation (12) must be non-zero.

For a proof of the realizability conditions outlined above see Wall [4].

Most transfer functions encountered are expandable by continued fractions. The fact that Type A expansion does not exist does not exclude the possibility of existance of a Type B expansion, and vice-versa. The sample problem illustrates this.

ILLUSTRATIVE EXAMPLE

The following example shows the development of a filter using the four techniques. The realizability conditions will first be checked, then the filter will be realized if possible. The function in question is

$$G(z) = \frac{1200\, z^4 + 320\, z^3 + 270\, z^2 + 45\, z + 1}{600\, z^4 + 100\, z^3 + 129\, z^2 + 10\, z} \tag{17}$$

The first step is to extract the constant term and normalize the leading coefficient in the denominator to unity. The result is

$$G(z) = 2 + \frac{.2\, z^3 + .02\, z^2 + .041666\, z + .0016666}{z^4 + .16666\, z^3 + .215\, z^2 + .016666\, z} \tag{18}$$

The various forms will now be examined.

For the Type 1A realization, matrix $\underline{M}$ is

$$\underline{M} = \begin{bmatrix} .2 & .02 & .041666 & .001666 & 0 & 0 & 0 & 0 \\ 1 & .1666 & .215 & .01666 & 0 & 0 & 0 & 0 \\ 0 & .2 & .02 & .041666 & .001666 & 0 & 0 & 0 \\ 0 & 1 & .1666 & .215 & .01666 & 0 & 0 & 0 \\ 0 & 0 & .2 & .02 & .04166 & .001666 & 0 & 0 \\ 0 & 0 & 1 & .1666 & .215 & .01666 & 0 & 0 \\ 0 & 0 & 0 & .2 & .02 & .041666 & .001666 & 0 \\ 0 & 0 & 0 & 1 & .1666 & .215 & .016666 & 0 \end{bmatrix} \tag{19}$$

Since b_0 is zero, we check minors D_1, D_2, D_3, .., D_7 of $\underline{M}$. The third minor D_3 is zero, thus the continued fraction expansion does not exist. If we blindly go ahead with the expansion, we get

$$G(z) = 2 + \cfrac{1}{5\, z + \cfrac{1}{3 + \cfrac{1}{4z^2 + \ldots}}} \tag{20}$$

The $4\,z^2$ term cannot be realized with the building blocks presented here.

To check the existence of the Type IB realization, we write

$$G(z) = 2 + \frac{1 + 25\,z + 12\,z^2 + 120\,z^3}{10\,z + 129\,z^2 + 100\,z^3 + 600\,z^4} \tag{21}$$

Eliminating the first term of the numerator, we find

$$G(z) = 3.21 + 0.1\,z^{-1} + \frac{-\,.2568166\,z^2 - .10166\,z^3 - 1.21\,z^4}{.01666\,z + .215\,z^2 + .166\,z^3 + z^4} \tag{22}$$

The matrix $\underline{M}$ corresponding to the above ratio is

$$\begin{bmatrix} -\,.2568166 & -\,.101666 & -\,1.21 & 0 & 0 & 0 \\ .01666 & .215 & .166 & 1 & 0 & 0 \\ 0 & -\,.2568166 & -\,.101666 & -1.21 & 0 & 0 \\ 0 & .0166 & .215 & .166 & 1 & 0 \\ 0 & 0 & -\,.2568166 & -\,.10166 & -1.21 & 0 \\ 0 & 0 & .0166 & .215 & .166 & 1 \end{bmatrix} \tag{23}$$

All six minors of this matrix are non-zero, hence the expansion exists. The final expansion is

$$G(z) = 3.21 + .1\,z^{-1} + \cfrac{1}{-\,.064879\,z^{-1} + \cfrac{1}{-\,1.2323 + \cfrac{1}{-\,.01197 + \cfrac{1}{-\,16.946z^{-1} + \cfrac{1}{.03427}}}}}$$

The circuit is simply one realizing $3.21 + .1\,z^{-1}$ placed in parallel with the Type IB circuit for the expansion shown above.

To check the existence of a Type II A realization, we check the odd-numbered principal minors of $\underline{M}$ in Equation 19. The third minor is zero, thus the II A realization is non-existent.

The conditions necessary for a Type II B realization are met by matrix $\underline{M}$ in Equation (23). We therefore expand, resulting in the function

$$G(z) = 3.21 + .1\,z^{-1} + \cfrac{1}{-.0649\,z^{-1} - .8115 + \cfrac{1}{-45.5311\,z^{-1} + 128.14 + \cfrac{1}{-.0016z^{-1} - .0051}}} \tag{24}$$

The constant terms are all non-zero, so this is a realizable function.

An interesting special case arises when G(z) is the ratio of an odd to an even polynomial (or even to odd). For example if G(z) is an odd over an even polynomial, Type II A realization will be of the form

$$G(z) = \cfrac{1}{B_1 z + \cfrac{1}{B_2 z + \cfrac{1}{\ddots + \cfrac{1}{B_n z}}}} \tag{26}$$

Similarly IIB yields the form

$$G(z) = \cfrac{1}{B_1 z^{-1} + \cfrac{1}{B_2 z^{-1} + \cfrac{1}{\ddots + \cfrac{1}{B_n z^{-1}}}}} \tag{27}$$

If G(z) is even over odd, the same forms naturally evolve.

CONCLUSION

Four techniques for developing realizations of digital filter transfer functions using continued fraction expansion have been presented. Each realization has corresponding realizability conditions which must be met in order to perform the expansion. All the resulting filters are canonic with respect to delays, and all except one (Type IIB) are canonic with respect to multipliers. Type IA, IB, and IIA realization use 2n multipliers for a general nth order system, while the Type IIB realization uses 2n + 1.

One circuit form for each of the four expansions has been shown. It is likely that other circuits can be devised using the same expansions shown in this work. These results are not limited to the digital domain; the results can equally well be applied to the continuous domain for uses such as realizing transfer functions on a analog computer.

REFERENCES

1. R. Edwards, J. Bradley and J. B. Knowles, "Comparison of noise performances of programming methods in the realization of digital filters," Proc. Symp. on Computer Processing in Communications, Vol. XIX, PIB-MRI Symposia Series, 1969, pp. 295-311.

2. O. Herrmann and W. Schüssler, "On the accuracy problem in the design of nonrecursive digital filters", Archiv der Elektrischen Ubertragung, vol. 24, 1970, pp. 525-526.

3. E. Avenhaus and W. Schüssler, "On the approximation problem in the design of digital filters with limited word lengths," Archiv der Elektrischen Ubertragung, vol. 24, 1970, pp. 571-572.

4. H. S. Wall, "Analytic Theory of Continued Fractions", Van Nostrand, 1948.

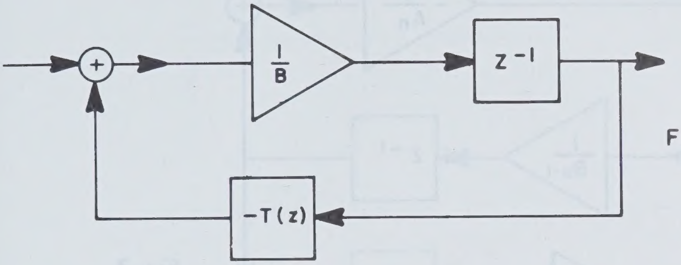
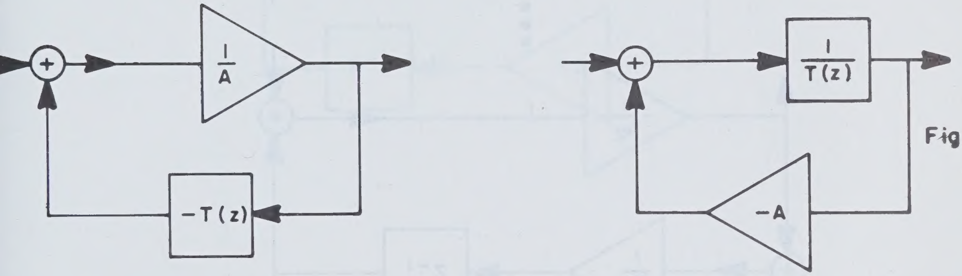


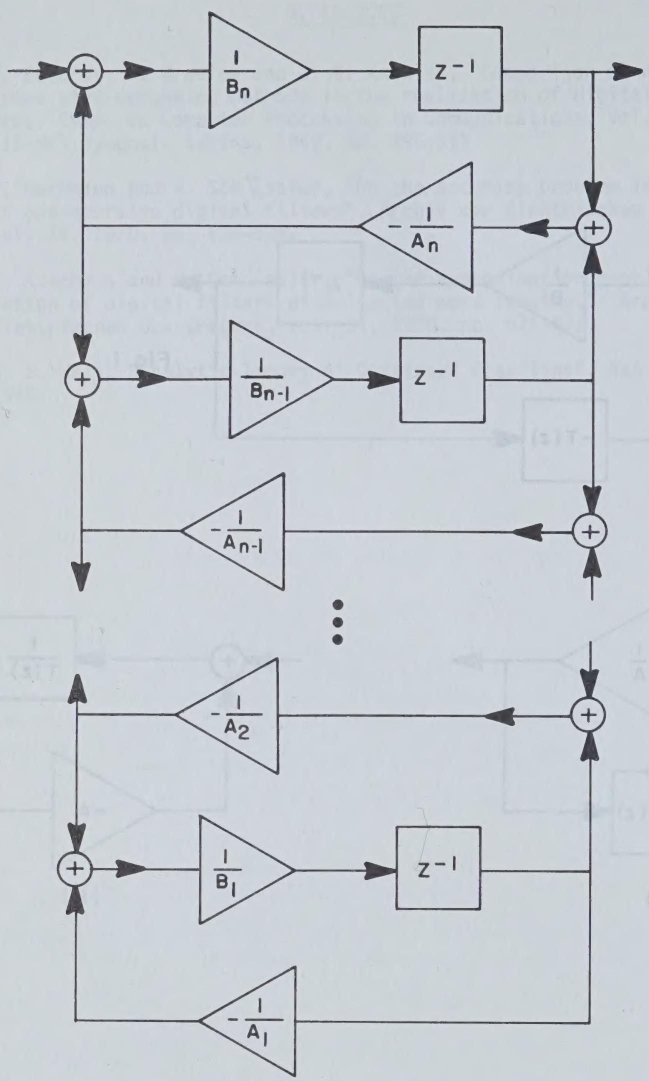
Fig. I



(A)

(B)

Fig. 2



1/B_n
z^-1
-1/A_n
1/B_n-1
z^-1
-1/A_n-1
-1/A_2
1/B_1
z^-1
-1/A_1

Fig. 3

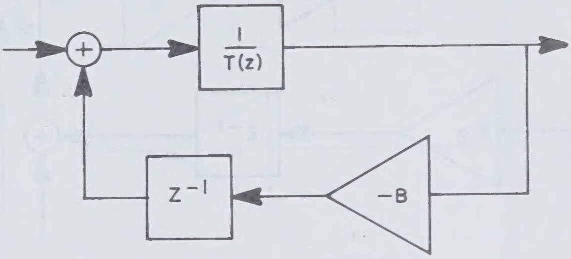


Fig. 4

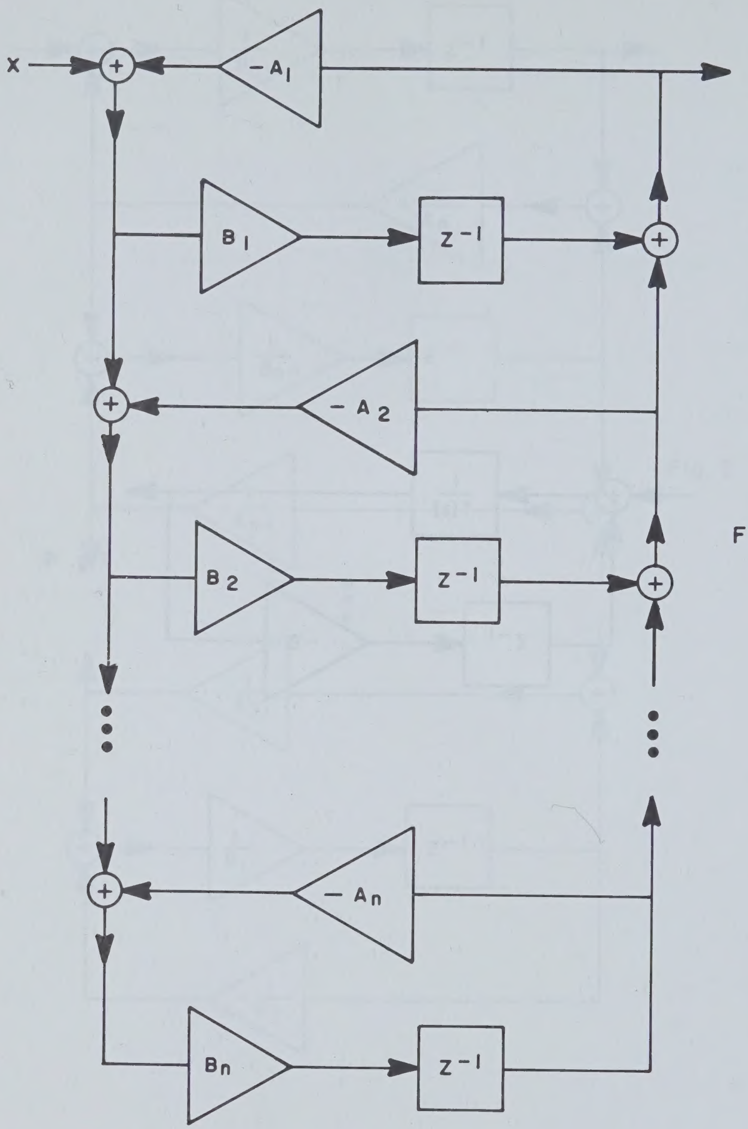

X
$-A_1$
B_1
z^{-1}
$-A_2$
B_2
z^{-1}
$-A_n$
B_n
z^{-1}

Fig. 5

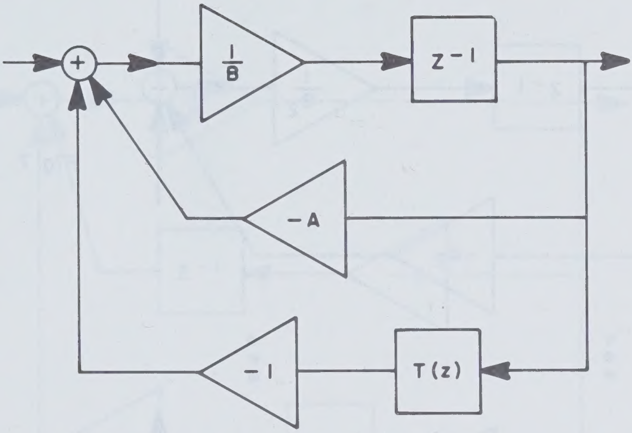
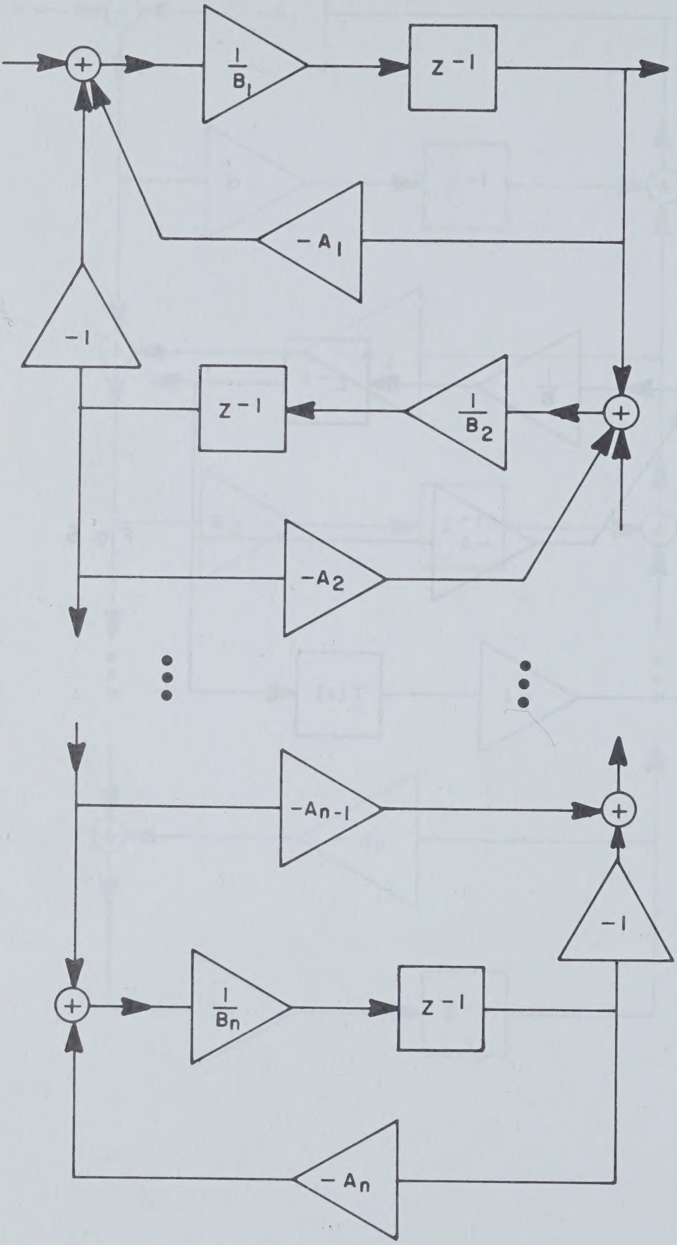


Fig. 6



1/B1
Z -1
-A1
-1
Z -1
1/B2
-A2
-An-1
-1
1/Bn
Z -1
- An

Fig 7

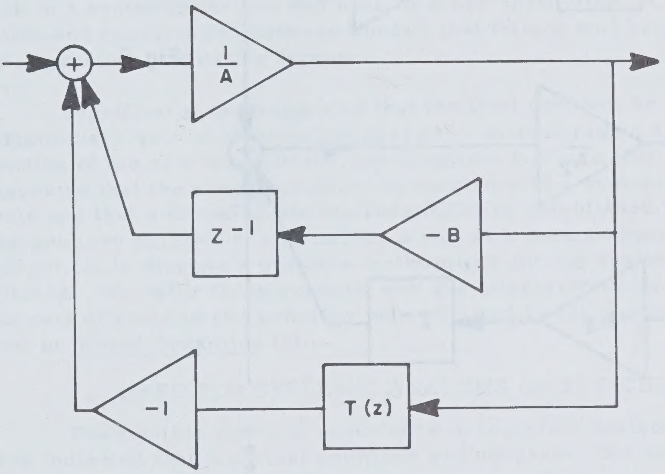


Fig. 8

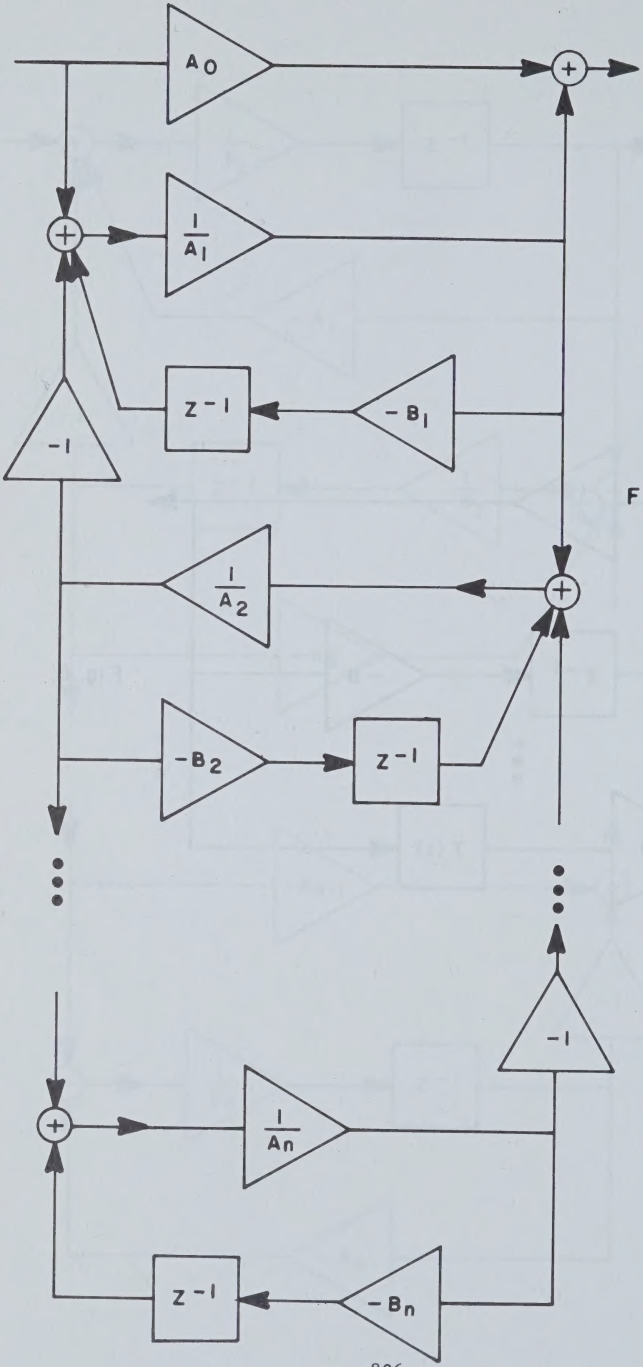

A0
1/A1
-1
Z-1
-B1
1/A2
-B2
Z-1
-1
1/An
Z-1
-Bn

Fig. 9

DYNAMICS OF LESS DEVELOPED COUNTRIES

KAN CHEN and H. E. HOELSCHER
School of Engineering
University of Pittsburgh
Pittsburgh, Pennsylvania 15213

INTRODUCTION: TWO ISSUES

Our world faces two problems having global significance and magnitude. Both are connected with the so-called less developed countries (LDC's). Three quarters of the world population in the LDC's are striving to improve their living conditions but are not making nearly enough progress; at the same time industrialization, recognized as the only possible solution to the first problem, is suspected to pose a fundamental threat to the total environment of our world in the long-range future. In a recent paper (1), we have suggested that these two problems represent two different, though interrelated, issues. They should be attacked separately, each in a systemic fashion and each by a new institution. Close communication and coordination between the two institutions will help to reconcile the apparently conflicting issues.

Specifically, we suggested that the first problem be tackled with a systemic analysis of efficacy and that such an analysis be a continuing function of the new International Development Institute (IDI). We further suggested that the second problem be tackled with a systemic analysis of goals and that a Global Goals Institute (GGI) be established for that task. The primary purpose of this paper, along with its accompanying papers (2, 3, 4), is to discuss a possible methodology for the system's analysis of efficacy. We refer those readers who are interested in the systemic analysis of goals to our previous paper (1) and to Jay Forrester's recent book on World Dynamics (5).

NEED FOR SYSTEMIC ANALYSIS OF EFFICIENCY

Twenty-five years of experience in technical assistance to LDC's have indicated that technical solutions are necessary but not sufficient to assure the efficacy of development programs. Machines must be maintained to function properly; goods produced by the machines must be distributed, etc. After its initial introduction to the LDC's, technology can be propagated only if the relevant activities in the countries are reorganized to utilize the technology, and if the new organizations are manned by people who motivate themselves and behave differently than they did previously. Thus, if the role of science and technology is considered only to provide solutions to technical problems, the solutions will be piecemeal and their contributions to the development process will be only marginal. The more central role of science and technology is that of a powerful change agent -- an agent for changing the social organization in such a way that industrialization systems will be accepted, take root and propagate. Kenneth Boulding pointed out that "the difference between a rich and a poor country lies mainly in the level of organization which it has attained."(6) There is a need for understanding, and enhancing the role of science and technology in social reorganization for development.

What has been said about technical solutions can also be said about "economic solutions" -- policy decisions for development based only on considerations of economic variables. The need for considering the non-economic variables and their systemic interrelationship with one another in a development process is now recognized by an increasing number of workers in the field of international development. However, the causal relationship can only be described qualitatively -- and even the qualitative relationship has not been tested in policy decisions. (7) Quantitatively, the United Nations has collected and analyzed a number of non-economic indicators related to development. However, the analysis can show only correlational rather than causal relationships. (8) What is needed, then, is a systemic analysis which will indicate quantitative causal relationships among all significant variables relevant to development -- a model which can help predict changes in the state of development (and its impact on the environment) predicated on various assumptions of exogeneous and decision variables. Such a dynamic model would be an effective tool for policy planning in the LDC's.

A POSSIBLE METHODOLOGY FOR MODELING LDC DYNAMICS

Any attack on the problem of development in the required systemic terms must affect the system with a time constant less than the time constants for the deteriorative processes which are at work. Furthermore, any such attack must recognize fully the systems character of the development process and the many interrelated and interacting threads which make up the fabric of a nation's welfare. We suggest that one process meets these criteria, namely the process of industrialization leading to a viable and productive industrial system for a nation. Such a system can be a primal force, feeding the economic, the social, the political and the cultural systems of the LDC's while recognizing them and involving them in the process.

If this argument is acceptable, then a possible methodology for modeling LDC dynamics (not for the purpose of explaining what has failed to take place, but for the purpose of generating and evaluating effective policies) would be to begin with the modeling of single industries, then the modeling of the interlocking relationships among the key industries, and finally the modeling of the interaction between the industrial system and the entire society of the LDC. In other words, three "Levels" of models are suggested, each increasing in difficulty and uncertainty over the other. Thus:

The "Level-1" models describe single industries. A typical model at this level is shown in Figure 1. (2) Such a model may be generated from an identification of the industrial process steps necessary within the industry and from the material, energy, and monetary balances that characterize the technology used by the industry. Of peripheral interest to the purpose of this paper, but significant to the systemic analysis of goals, is the capability of the Level-1 models to predict the kind and amount of wastes and pollutants generated by the industrial process and the cost of removing them.

The "Level-2" model describes a system of interacting industries operating in a given region of a nation or the entire nation. As shown in Figure 2, an example system includes linkages among individual industries and the parameters associated with these linkages are determined from field data and from the analysis of Level-1 models. The dynamics of the Level-2 model is discussed in an accompanying paper.

The "Level-3" model includes the principal interaction of the industrial system with the social system. The dynamics of the Level-3 model will be the subject of discussion in the remainder of this paper.

THE "LEVEL-3" MODEL

Level-3 modeling is at once most crucial and most difficult. It is most crucial because it espouses the key variables and their interrelationships in the whole development process. When it is adequately established, the Level-3 model will provide the anchor for the Level-1 and Level-2 models in development planning. However, Level-3 modeling is also most difficult, for the complex and comprehensive interactions among the cultural, social, political, economic, as well as technological forces in the development process are still poorly understood. Quantitative economic growth models exist, but they tend to focus on capital accumulation, which is only one of many aspects and is not the most fundamental aspect in the total development process. (9) Some recent books have provided excellent and comprehensive reviews of the literature on development processes (7, 10, 11), but understandably they can only suggest new qualitative theories or new directions for research on a speculative basis. Meanwhile the world continues to turn and the development gap between the LDC's and the rest of the world continues to widen. As we cannot stop the clock to wait for research in social science to catch up, we propose an action-oriented research procedure for the Level-3 modeling.

Basically the procedure calls for the construction of an approximate model, quantitative to the extent feasible, based on the best, though imperfect, understanding we have on the total development process. As Forrester puts it, "man acts at all times on the models he has available. Mental images are models. We are now using those mental models as a basis for action."(5) Next, the approximate model will be subjected to sensitivity analysis. The sensitive parts of the model, its sensitive structure as well as its sensitive parameters, will suggest a program for field work, empirical research, and for socio-technological experimentation so that the uncertainties about the sensitive parts can be reduced. The feed-back will be used to improve or correct the original approximate model. This will be a continuing process of learning and updating. Meanwhile the latest Level-3 model, once it passes an initial testing for reasonableness, will always be available for policy generation and evaluation.

Specific Example

To illustrate the first step in the proposed action-oriented, research procedure, an approximate quantitative model has been constructed to describe the development process on the basis of Rostow's "take-off" theory (12). The nonlinear feedback system in Figure 3 shows the admittedly over-simplified dynamic system of a LDC. Rostow defines take-off as requiring all three of the following related conditions (paraphrased for the current purpose):

(1) A rise in the rate of productive investment above a threshold
(2) The development of a growing manufacturing sector (leading sector)
(3) The existence of a political, social and institutional framework that permits the growth at a compound rate.

These three conditions are represented, respectively, by the threshold, the industrial system, and the social system in Figure 3. The industrial system may also be considered a simplified version of the Level-2 model. (4)

It will be interesting to see how the model in Figure 3 could be used for explaining and testing policies for take-off. Beginning with the left-hand side of the diagram, we see the corrective action as the sum of an external stimulus (such as foreign aid) and self-generating efforts. The corrective action is to overcome the deteriorative force in order to lift industrial output above the subsistence level. The nonlinearity in the feedback path consists of the threshold for take-off. Below the threshold, the industrial output would be used only for consumption at the subsistence level. There will be capital accumulation only if the industrial output exceeds the threshold. However, the capital cumulation will not be put to productive reinvestment (i. e., the surplus will flow into the hands of those who will sterilize it by hoarding, luxury consumption, or low-productivity use) unless there is a continuing and increasing supply of professional skills, entrepreneurs, etc., within an adequate institutional framework for development. The long time constant (one generation or 20 years) of the social system characterizes the relatively long process of social change in developing these skills and institutional framework.

It is suspected that the LDC, as represented by Figure 3, may or may not take-off, depending on how much and how long the industrial output exceeds the threshold. Although the system in Figure 3 appears extremely simple, it is practically impossible to depend solely on intuitive judgment as to the magnitude and duration of the external stimulus (foreign aid) that is necessary to cause the LDC to take off. Experimentation on computers is then necessary and helpful to evaluate the policy decision quantitatively.

Figure 4 shows a number of computer runs which would illustrate the potential power and significance of dynamic modeling.* Curve 1 shows that the industrial output would drift below the subsistence level in the absence of external stimulus. Curve 2 shows how the industrial output would climb above the threshold and eventually increase exponentially with time, as expected, if substantial external stimulus is applied and maintained indefinitely. Curves 3, 4 and 5 show the effects of cutting off external stimulus at the end of 2, 6 and 12 years, respectively. It is interesting to note that the industrial output reaches the threshold after only 1 year. However, because of the system dynamics, the effect of industrialization has not "taken root" at the end of the second year so that, if the external stimulus were removed then (Curve 3), the system would revert to below subsistence level. In fact, Curve 4 shows that the effect of industrialization has barely taken root only after 6 years so that slow recovery and eventual take-off would be possible if the external stimulus were removed then. From the computer curves in Figure 4, one could see the danger for foreign aid officials who, looking at the flat portion of industrial output from the 3rd to 5th years, might easily get frustrated and give up on the LDC's by recommending the cutoff or curtailment of foreign aid to the country too soon. In other words, the short-term growth or lack of growth of industrial output should not be used as the sole indicator of the LDC's capability of self-sustaining growth in the long run. A dynamic model of the whole development process could thus be an essential tool for policy decision.

Possible Extensions

The specific but highly simplified system in Figure 3 was constructed only for illustration and should not be taken too seriously except as a starting point. It does represent a possible first step towards Level-3 modeling in accordance with the action-oriented research procedure described previously. In the realistic modeling of a given LDC, the corrective and the deteriorative forces will have to be country-specific (overpopulation, shortage of skilled labor, etc.). Although the industrial system could be considered as driven by a set of final demands, supply-oriented model might be more appropriate to use in order to reflect the constraints imposed by limited supply factors in the LDC. (13) Perhaps the most important and most interesting extension will be the meaningful expansion of the transfer function representing the social system. The expansion may take advantage of the stage theories of development (in which case the mathematical representation of the social system might include switching changes or drastic nonlinearities to reflect structural transformation), the technological adaptation theory (14) (in which case the mathematical expression would be a diffusion model), the social learning theory (7, 11) (in which case a Markov chain could be the appropriate form of mathematical modeling), or a combination of all these theories. In any event, these possible extensions should be considered in the context of the given LDC, and a judicious expansion of Figure 3 should be implemented before the next steps in the procedure (sensitivity analysis field work, etc.) are implemented. The sensitivity analysis applied to the system in Figure 3, as given in an accompanying paper (3), is again for illustration and not the result of a realistic implementation of the proposed action-oriented research procedure.

Summary

The two issues underlying the dynamics of less developed countries, the issue of efficacy and the issue of goals, should be distinguished. This

---

* The authors are indebted to Mr. Koustubh Oka for conducting the computer experimentation on the model.

paper, concentrating on the first issue, presents a methodology for modeling the relevant dynamics at three levels. Level-1 models, restricted to single industries, can be successfully generated without significant difficulty. Level-2 models are generable, given the data, and provide important insights to interindustry relationships. Level-3 models, representing the entire development process, are most difficult to generate. However, for the practical purpose of generating and evaluating development process, an action-oriented research procedure can be used to construct useful Level-3 models which can be improved and updated on a continuing basis.

REFERENCES

(1) Chen, K. and H. E. Hoelscher, "The Role of Science and Technology in Less-Developed Countries -- A Systemic Analysis," Record of the Woods Hole Conference on the Role of Science and Technology in International Development in the 1970's, August, 1971.

(2) Hoelscher, H. E., Y. T. Shah and H. M. Beisel, "A Detailed Dynamic Analysis of a Soybean Extraction Plant"(In Preparation).

(3) Oka, K., "Experiments with a Dynamic Model of a Less Developed Country," 9th Annual Allerton Conference on Circuit and System Theory, October, 1971.

(4) Vogt, W. G., and M. H. Mickle, "A Dynamic Input-Output Model for an Economic System," 9th Annual Allerton Conference on Circuit and System Theory, October, 1971.

(5) Forrester, J. W., World Dynamics, (Wright-Allen Press), 1971.

(6) Boulding, K., "Some Questions on the Measurement and Evaluation of Organization," Beyond Economics, (University of Michigan Press), 1968.

(7) Kunkel, J. H., Society and Economic Growth, (Oxford University Press), 1970.

(8) McGranahan, D. V., "Analysis of Socio-Economic Development Through a System of Indicators," Annals of American Academy of Political and Social Sciences, Vol. 393, January, 1971.

(9) Ackley, G., Macroeconomic Theory, (Macmillan), 1961.

(10) Hetzler, S. A., Technological Growth and Social Change, (Praeger), 1969.

(11) Dunn, E. S., Jr., Economic and Social Development: A Process of Social Learning, (Johns Hopkins Press), 1971.

(12) Rostow, W. W., The Stages of Economic Growth: A Non-Communist Manifesto, (Cambridge University Press), 1960.

(13) Winger, A. R., "Supply Oriented Urban Economic Models," AIP Journal, Vol. 35, No. 1, 1969.

(14) Brown, L., Diffusion Dynamics, Royal University of Lund, Sweden, 1968.

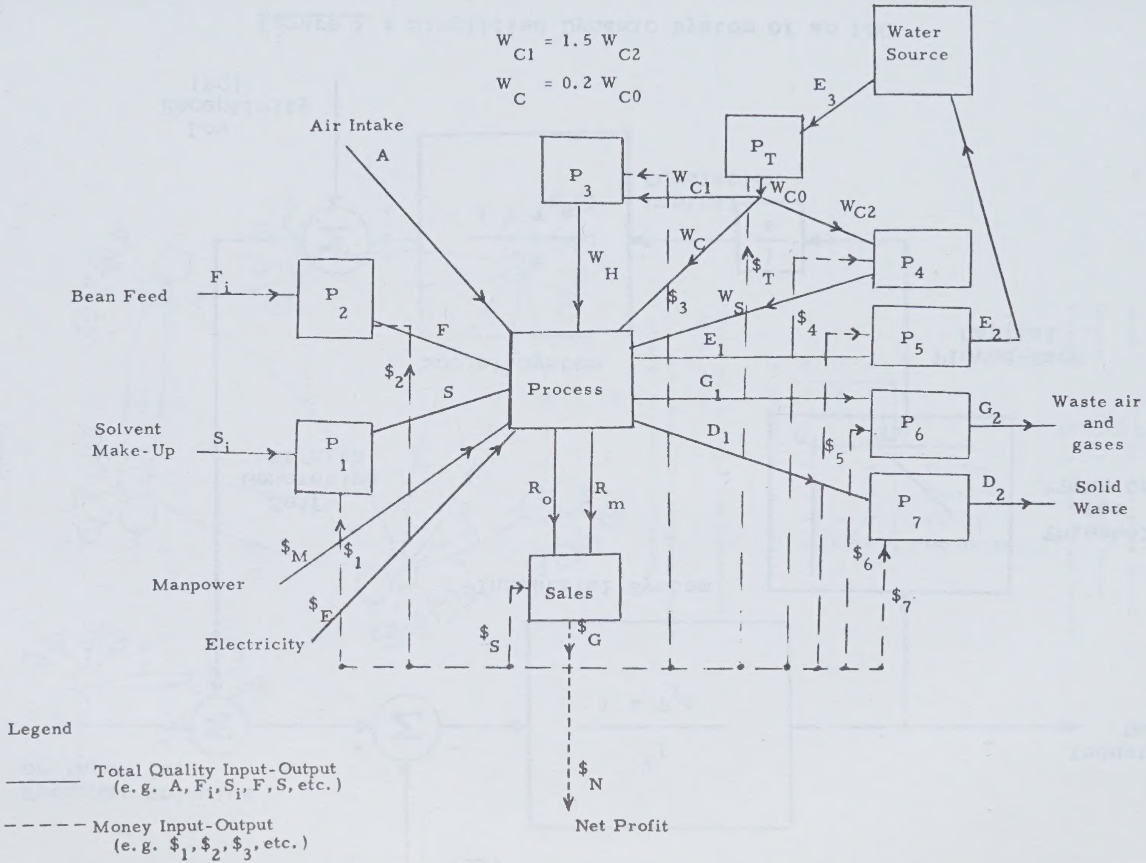

$W_{C1} = 1.5 W_{C2}$
$W_C = 0.2 W_{C0}$
Water Source
E_3
P_T
W_{C0}
W_{C1}
W_{C2}
W_C
P_3
W_H
Air Intake
A
Bean Feed
F_i
P_2
F
$\$_2$
S
Solvent Make-Up
S_i
P_1
$\$_1$
$\$_M$
Manpower
$\$_E$
Electricity
Process
$\$_3$
$\$_T$
W_S
P_4
$\$_4$
E_1
P_5
E_2
G_1
P_6
G_2
$\$_5$
Waste air and gases
D_1
P_7
D_2
Solid Waste
$\$_6$
$\$_7$
R_o
R_m
Sales
$\$_S$
$\$_G$
$\$_N$
Net Profit
Legend
Total Quality Input-Output (e.g. A, F_i, S_i, F, S, etc.)
Money Input-Output (e.g. $\$_1$, $\$_2$, $\$_3$, etc.)

Figure 1 - Analysis of Soybean Process

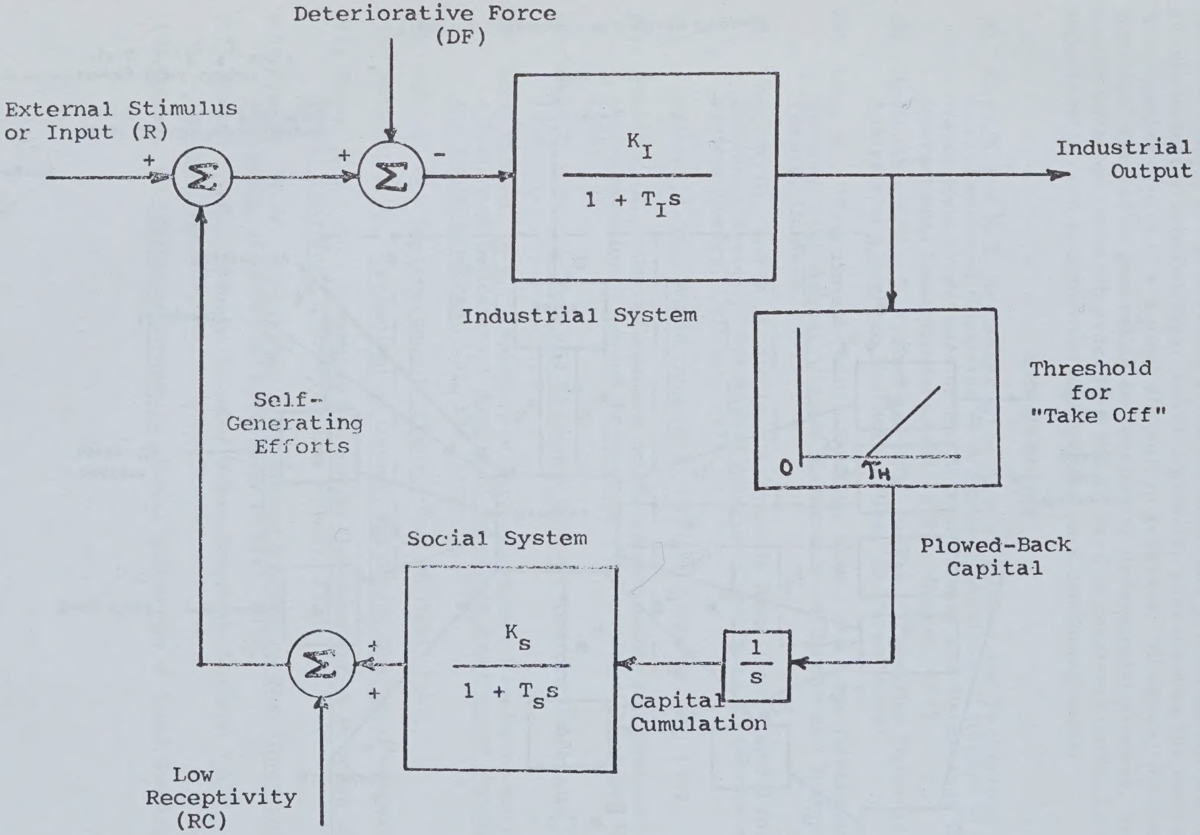

Deteriorative Force
(DF)
External Stimulus
or Input (R)
+
+
-
Σ
Σ
$\frac{K_I}{1 + T_I s}$
Industrial System
Industrial
Output
0
T_H
Threshold
for
"Take Off"
Plowed-Back
Capital
$\frac{1}{s}$
Capital
Cumulation
Social System
$\frac{K_S}{1 + T_S s}$
+
+
Σ
Self-
Generating
Efforts
Low
Receptivity
(RC)

Figure 3 : Simplified Dynamic System of an LDC

Figure 2

LEGEND

Circle :: ○ :: Personal Demand
Square :: □ :: Investors Demand
Octagon :: ⯃ :: Foreign Demand
Triangle :: △ :: Government Demand

	Processing			Personal	Investors	Foreign	Government
A	3	2	0	7	2	2	6
B	2	6	0	5	3	3	5
C	0	0	1	1	0	1	0

Table I Transactions Table

Three industry "types" are represented in this model.

The intra industry flows are given arbitrary values for purposes of illustration.

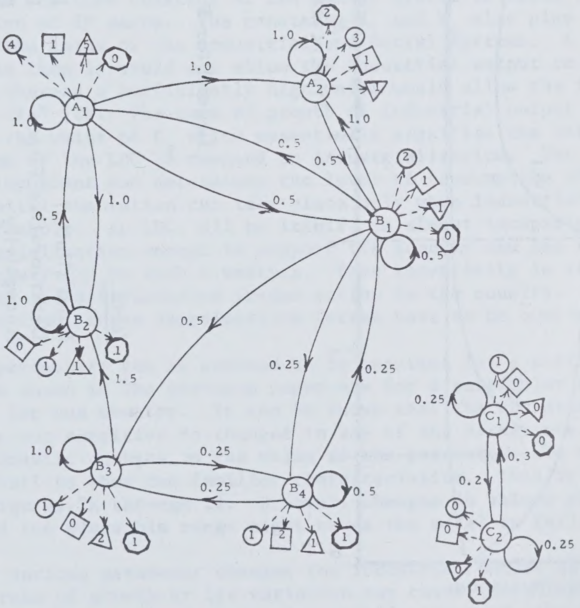
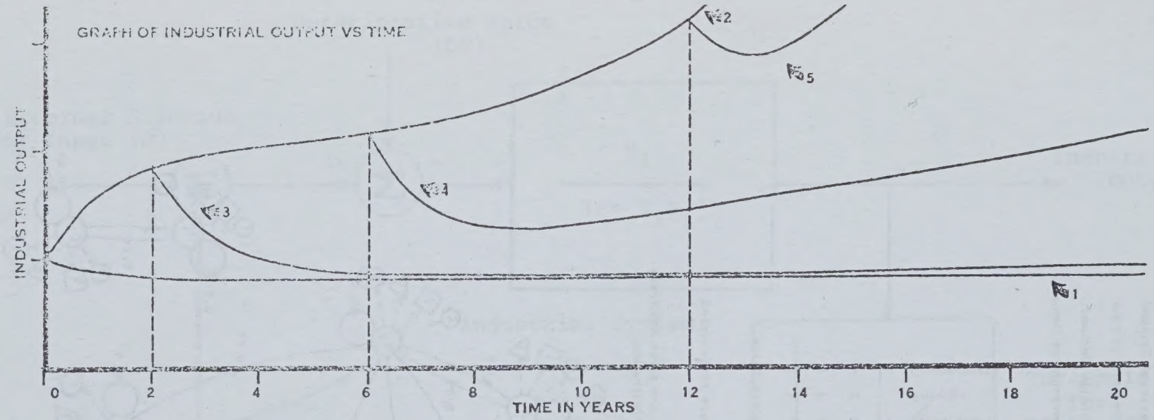

A1
A2
B1
B2
B3
B4
C1
C2
1.0
0.5
0.25
1.5
0.3
0.2
4
1
5
0
2
3

Figure 2

Figure 4



Legend for Figure 4

Item	Symbol
Input (Units)	R = 2.5
Deteriorative Force	DF = 0.6
Receptivity	RC = 0.3
Threshold	TH = 1
Industrial System	
Gain	$K_I = 1$
Time Constant (yrs.)	$T_I = 1$
Social System	
Gain	$K_S = 2$
Time Constant (yrs.)	$T_S = 20$

EXPERIMENTS WITH A DYNAMIC MODEL OF A LESS DEVELOPED COUNTRY

KOUSTUBH OKA
University of Pittsburgh
Pittsburgh, Pennsylvania.

ABSTRACT

An Analog Computer model of a simple Industrial System of a less developed country (LDC) is presented. System sensitivity to parameter variation is discussed and industrial output is studied as a function of time. As another approach, the effect of fixed capital over a period of time (such that the product of capital and time is constant) on industrial output is studied over a period of 25 years.

INTRODUCTION

The purpose of this paper is to elaborate on the validity of the Analog Computer model of Figure 1 [1]. It will discuss the parameters of the model, its sensitivity to changes in these parameters and a new and different problem that can be studied by using this model.

THE MODEL AND ITS SENSITIVITY

In the simplified dynamic system of the LDC (Figure 1) only a few but significant parameters of the real world system are taken into consideration. Essentially the time constant for the process of industrialization must be small as compared to that of the process of social change. The ratio is 1:10 and the time constant of the Social System is taken to be about a generation or 20 years. The constants K_I and K_S also play an important role being the gains of the industrial and social systems. A very low value of K_I (less than 1) would not allow the industrial output to even reach threshold whereas a sufficiently high value would allow the industrial output to "take off" [2]. The rate of growth of industrial output on "take off" depends on the value of K_S which essentially signifies the ability of the social system of the LDC to respond to industrialization. The threshold itself is very important and determines the level of consumption of a country, and capital cumulation can take place only when industrial output exceeds the threshold. An LDC, all by itself, is almost incapable of achieving industrialization enough to support its economy and has an inherently low receptivity to such a venture. This receptivity is significantly lower than the deteriorative forces active in the country. For any economic success then, these deteriorative forces have to be overcome.

All these parameters can be assumed to be constant for a particular LDC. The curves shown in the previous paper are for a particular set of parameters i.e. for one country. It can be shown that the characteristics of the model are not sensitive to changes in any of the parameters. In fact even if fairly drastic changes in the value of the parameters are made, the model does not fail to show the familiar characteristics. This is verified by looking at Figures 2A through 2E. However, changes in values of parameters to beyond the plausible range might cause the model to fail.

Because of various parameter changes the industrial output of an LDC may change its rate of growth or its variation may cause the threshold to be reached in various periods of time. Essentially, however, the dynamic behaviour of the model is the same as the expected behaviour of a real world

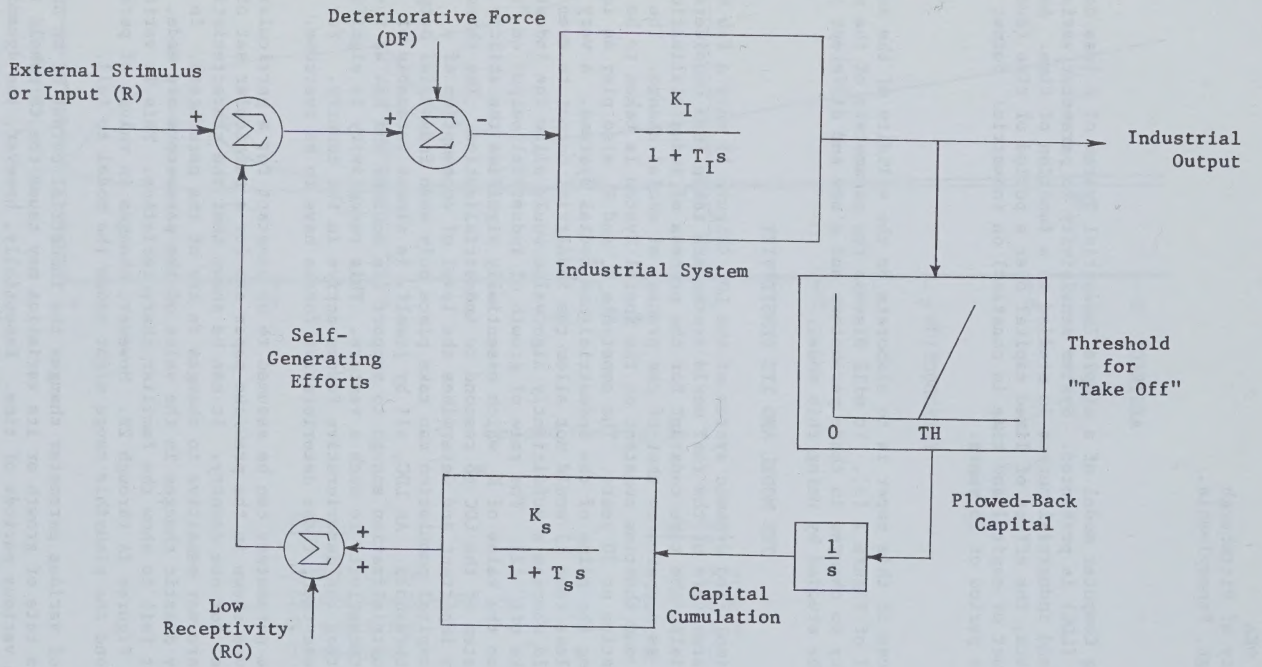

Deteriorative Force
(DF)
External Stimulus
or Input (R)
+
+
-
$\frac{K_I}{1 + T_I s}$
Industrial System
Industrial
Output
Threshold
for
"Take Off"
0
TH
Plowed-Back
Capital
$\frac{1}{s}$
Capital
Cumulation
$\frac{K_s}{1 + T_s s}$
+
+
Self-
Generating
Efforts
Low
Receptivity
(RC)

Figure 1 : Simplified Dynamic System of an LDC

system subjected to the same set of parameter values.

In each of the Figures 2A through 2E : curve 1 shows how the industrial output drifts below subsistence level in the absence of external stimulus (input-R). Curve 2 shows how the industrial output would increase above the threshold and take off exponentially if substantial external stimulus is applied for an indefinite period of time. Curves 3, 4, 5 (and sometimes 6) show the effects of cutting off this input at various times. The purpose of the graphs is to show that the characteristics of the system do not change appreciably in shape and behaviour for changes in parameter values; hence the cut-off points are not precisely at particular moments in time, but rather approximately so. Parameter values for Figures 2A through 2E are summarized in Table 1.

Table 1

Summary of Parameter Values Used for Figures 2A through 2E

Figure	R	RC	DF	TH	Industrial System		Social System	
					K_I	T_I	K_s	T_s
2A	2	0.3	0.6	1.0	1.0	1.0	2.0	20.0
2B	2	0.3	0.6	1.0	2.0	1.0	1.0	20.0
2C	2	0.3	0.6	1.0	2.0	1.0	0.5	20.0
2D	2.5	0.3	0.6	1.0	1.0	1.0	2.0	20.0
2E	2	0.3	0.6	0.75	1.0	1.0	2.0	20.0

R is the input (external stimulus)
RC is the receptivity
DF is the deteriorative force
TH is the threshold
K_I is the gain of the industrial system
T_I is the time constant of the industrial system
K_s is the gain of the social system
T_s is the time constant of the social system

A DIFFERENT PROBLEM

As another problem it may be considered that a certain amount of capital "C" is available for investment in an LDC, to serve as its external stimulus. Now this entire amount C can serve as input for say a period of 1 year or it could be divided up into 0.5C as input for 2 years or 0.25C for 4 years or 0.04C for 25 years. (Note that for the sake of simplification the interest on the amount that would be due is not taken into consideration). Thus the input is a pulse of varying heights, but the area under the curve is constant. Figure 4 shows the industrial output curves. Curves 3, 4, and 5 show the effects of supplying little external stimulus for long periods of time - ultimate drift to below the subsistence levels. Curves 1 and 2 are the effects of applying large external stimulus for short periods causing the system to "take off". The curves then reflect the theory that if a lump sum amount was invested in an average LDC for a short period of time, Rostow's "take off" would indeed occur and aid in the development of the LDC. On the other hand if this investment was disbursed over a longer period of time with smaller cuts per unit of time, the investment would be consumed and the process would drift to below subsistence level.

Table 2 gives the values of the total industrial output as a function of external stimulus and time. Note the difference between the values of the integrated output for a pulse applied for one year and half the pulse applied for two years. This suggests that the entire amount C invested for one year (on our scale) would be much more productive than half the amount invested over two years. The results, as before, seem to be counter-intuitive. Figure 3 shows the plot of Total Industrial Output vs Time.

Table 2

Total Industrial Output as a Function
of Time and External Stimulus

External Stimulus	Length of Time (Years)	Total Industrial Output
6.0	1.0	15.94
3.0	2.0	6.72
2.35	2.55	2.94
1.5	4.0	0.0
0.6	10.0	0.06
0.24	25.0	0.0

Once again the dynamic model of an industrialization process has been shown to be a handy tool for solving new problems and for looking at newer approaches to old ones.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to Dr. W.G. Vogt and Dr. M.H. Mickle for the guidance and encouragement they offered during the course of my work. I would also like to thank Dr. Kan Chen for introducing me to the subject and guiding me in the initial stages. I am very grateful to Dean H.E. Hoelscher for having given me the opportunity to work on this project.

REFERENCES

1. Chen, K., Hoelscher, H.E., "Dynamics of Less Developed Countries," 9th Annual Allerton Conference on Circuit and System Theory, October 1971.

2. Rostow, W.W., The Stages of Economic Growth : A Non-Communist Manifesto, (Cambridge University Press), 1960.

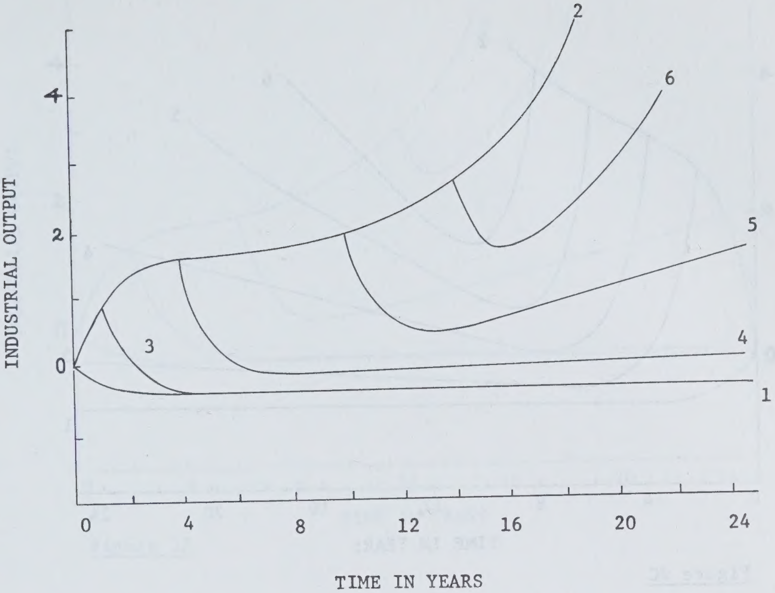


Figure 2A

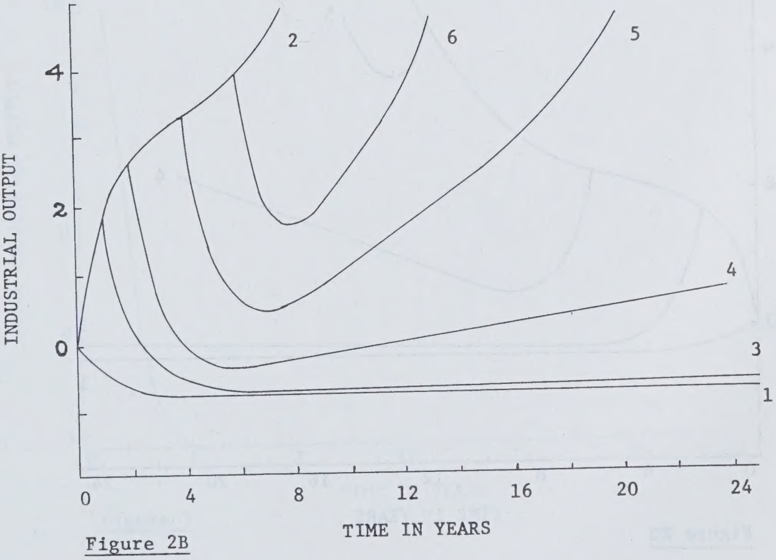


Figure 2B

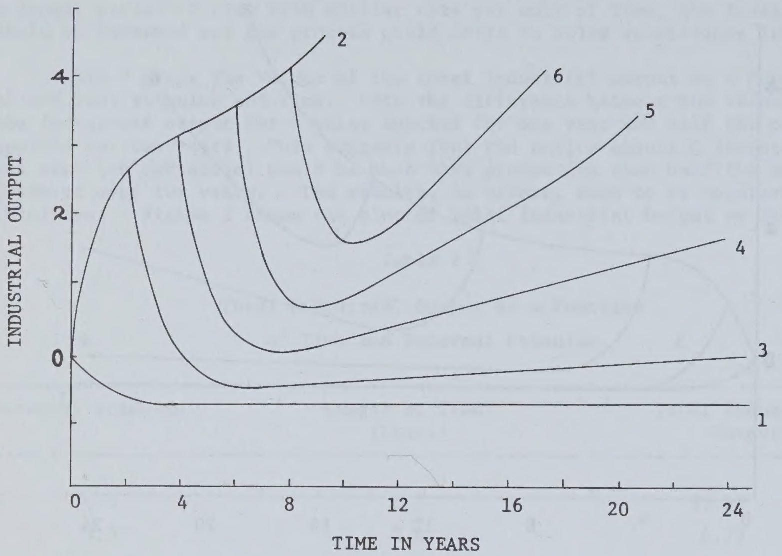


Figure 2C

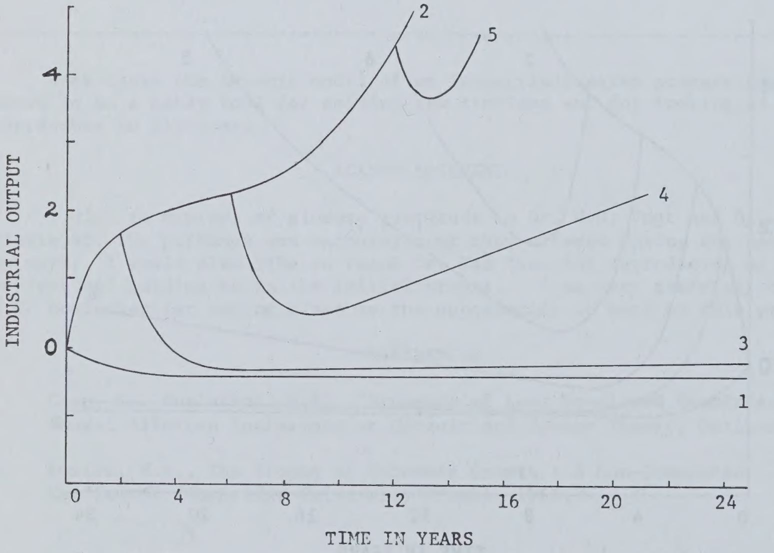


Figure 2D

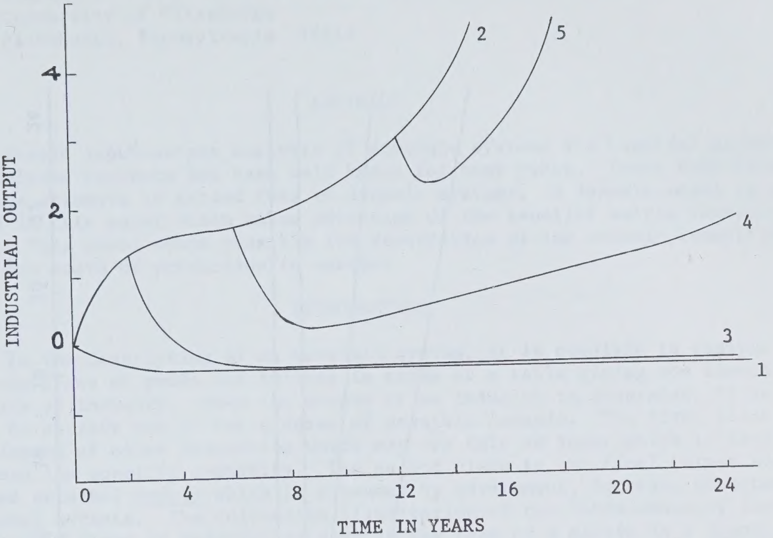


Figure 2E

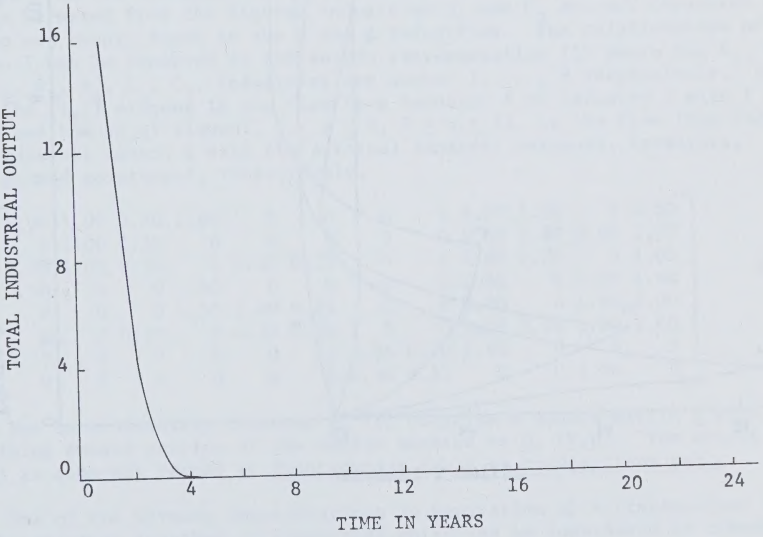


Figure 3

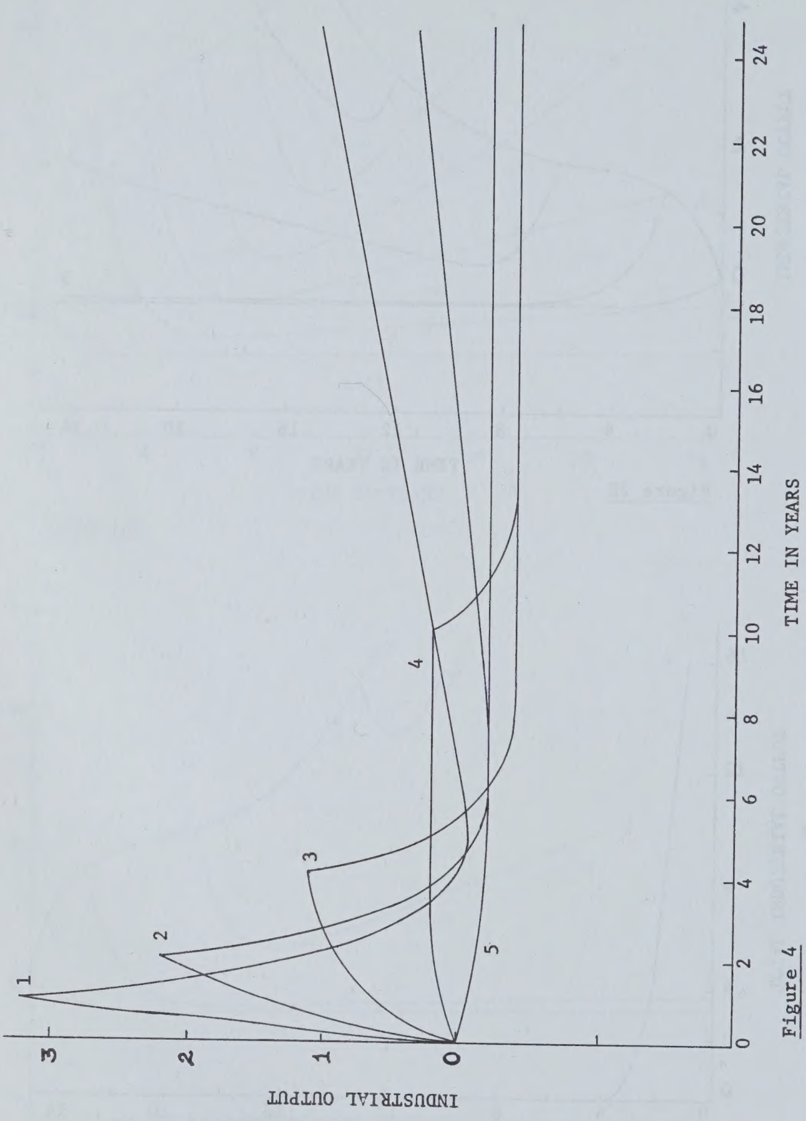

INDUSTRIAL OUTPUT
3
2
1
0
TIME IN YEARS
0
2
4
6
8
10
12
14
16
18
20
22
24
1
2
3
4
5

Figure 4

A DYNAMIC INPUT-OUTPUT MODEL FOR AN ECONOMIC SYSTEM

WILLIAM G. VOGT, MARLIN H. MICKLE AND HRAIR ALDERMESHIAN
University of Pittsburgh
Pittsburgh, Pennsylvania 15213

ABSTRACT

Static input-output analysis of economic systems via Leontief matrices and related concepts has been well known for many years. There have been various attempts to extend this to dynamic systems. A dynamic model is proposed in this paper which takes advantage of the Leontief matrix formulation. This model makes possible the description of the dynamic transition from one state of production to another.

INTRODUCTION

In the description of an economic system, it is possible to analyze the various flows of goods and dollars in terms of a table giving the inputs and outputs of industry. When the output of an industry is generated, it is used to satisfy one of two classes of possible demands. The first class is the demand of other industries which may use this as input which is used to produce its specific commodity. The second class is the final output here called external supply which is consumed by government, foreign, investor or personal markets. The collective illustration of the intra-industry input and output flows of material or cash in the form of a matrix is a transactions table. In economic terms the resulting model is an open one, that is, one which supplies external demands and does not take into account labor.

Consider the industrial complex of Figure 1 with industries A_1, A_2, B_1, B_2, B_3, B_4, C_1, C_2. Eight distinct industries are given with inputs and outputs from and to other industries and outputs to supply external demand. As can be noted from the figure, industries C_1 and C_2 are not dependent on and do not supply input to the A and B industries. The relationships of Figure 1 can be observed in the matrix representation (1) where the A_1, A_2, B_1, B_2, B_3, B_4, C_1, C_2, industries are number 1, ..., 8 respectively. In (1), the (i,j) element is the flow from industry i to industry j with $1 \leq i, j \leq 8$ and the (p,q) element, $1 \leq p \leq 8$, $9 \leq q \leq 12$, is the flow from industry p to external demand q with the external demands; personal, investors, foreign and government, respectively.

$$\begin{bmatrix} 1.00 & 1.00 & 0.50 & 1.00 & 0 & 0 & 0 & 0 & 4.00 & 1.00 & 0 & 5.00 \\ 0 & 1.00 & 0.50 & 0 & 0 & 0 & 0 & 0 & 3.00 & 1.00 & 2.00 & 1.00 \\ 0.50 & 1.00 & 0.50 & 0 & 0.50 & 0.25 & 0 & 0 & 2.00 & 1.00 & 0 & 1.00 \\ 0.50 & 0 & 0 & 1.00 & 0 & 0 & 0 & 0 & 1.00 & 0 & 1.00 & 1.00 \\ 0 & 0 & 0 & 1.50 & 1.00 & 0.25 & 0 & 0 & 1.00 & 0 & 1.00 & 2.00 \\ 0 & 0 & 0.25 & 0 & 0.25 & 0.50 & 0 & 0 & 1.00 & 2.00 & 1.00 & 1.00 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0.20 & 1.00 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.30 & 0.25 & 0 & 0 & 1.00 & 0 \end{bmatrix} \qquad (1)$$

The inter-industry elements of (1) comprise a square matrix $\underline{V}$ with the remaining demand portion of the matrix denoted by $\underline{D}$, $(\underline{V},\underline{D})$. The matrix $(\underline{V},\underline{D})$ is commonly termed an input-output table or transactions table.

One of the primary considerations in generation of a transactions table is the grouping together of industries which can be considered as common for the type of analysis for which the table is intended. The grouping together of industries into one unit is termed *aggregation*. On the other hand, it may be necessary to divide classical groupings to arrive at table entries giving rise to the term *disaggregation*. The table of individual flows comprise what is termed a completely *disaggregated system*.

While much of the analysis may be concerned with the aggregated system, many of the individual alternatives may not be distinguishable after aggregation. Therefore it will be desirable to have a means of conveniently aggregating the system. The inter-industry portion of the completely disaggregated transactions table has been denoted by the nxn square matrix $\underline{V}$. The aggregated matrix will then be a square matrix $\underline{U}$ with dimension m less than that of $\underline{V}$. Aggregation can be expressed as shown in (2) where $\underline{G}$ is an nxm matrix and G' its mxn transpose.

$$\underline{U} = \underline{G}'\underline{V}\underline{G} \tag{2}$$

Consider a numbering system for the industries of $\underline{V}$, $p=1,2,\ldots,N$ where the industries to be aggregated in each group have number $n_1, n_2-n_1, n_3-n_2, \ldots, n_q-n_{q-1}$ with $n_1<n_2<\ldots<n_q$, and $n_0 \equiv 0$. The general element of $\underline{U}$ is given by

$$u_{ij} = \sum_{p=n_{j-1}+1}^{n_j} \sum_{k=n_{i-1}+1}^{n_i} v_{kp} \tag{3}$$

From (3), the *aggregation matrix*, $\underline{G}$, has general element

$$g_{ij} = \begin{cases} 1 & : \; n_{j-1} \le i \le n_j \\ 0 & : \; \text{otherwise} \end{cases} \tag{4}$$

The supply matrix, $\underline{D}$, of $(\underline{V},\underline{D})$ is aggregated with matrix $\underline{G}$ by $\underline{G}'\underline{D}=\underline{E}$ where $\underline{E}$ is the demand portion of the reduced transactions table. Thus,

$$(\underline{U},\underline{E}) = (\underline{G}'\underline{V}\underline{G},\ \underline{G}'\underline{D}) \tag{5}$$

In a study of industrial development of developing countries it is desirable to choose a particular aggregation.[3] Industries which can operate at world levels of technology in their production process are termed Type A. Type B industries operate at less than world levels of technology and act solely as feeders to Type A industries. Type C industries operate at less than world levels of technology and exist as more or less independent units within the community. The results of the present paper apply to a general economic system. However the A, B, C type aggregation will be used for purposes of illustration and demonstration without any loss of generality.

The $(\underline{V},\underline{D})$ matrix of (1) will now be aggregated by the A, B, and C concept. The transpose of the aggregation matrix $\underline{G}$ for the example at hand is shown below in (6).

$$\underline{G}' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{6}$$

The aggregated transaction table, $(\underline{U},\underline{E})$, for (1) is thus given by $(\underline{G}'\underline{V}\underline{G},\underline{G}'\underline{D})$ and is shown in Table 1.

		Processing								
		A	B	C	Personal	Investors	Foreign	Government	Total supply for external demand (s_i)	Total production (x_i)
Output	A	3	2	0	7	2	2	6	17	22
	B	2	6	0	5	3	3	5	16	24
	C	0	0	1	1	0	1	0	2	3

Table 1. Transactions Table

The aggregation which has been performed is similar in concept to that by A.S. Manne and A. Rudra[7] for the economy of India where type A is the metals and machinery sector of the economy, type B is the universal intermediates, and type C is the food and fiber sector.

Input-Output tables of many developed and underdeveloped countries have been prepared to assist in economic analysis and economic forecasting. In America, for example, Input-Output tables of the U.S. Economy are prepared by the U.S. Department of Commerce, Office of Business Economics and are published in "Survey of Current Business." Input-Output tables of Greece are compiled at the Center of Planning and Economic Research. Among the other countries which have published detailed Input-Output tables are Israel, Egypt, Spain and Argentina.

Most of these Input-Output tables are compiled by government agencies and in a few cases by economic research institutions. Usually a delay of five to six years is incurred due to the tremendous amount of work involved in compilation of economic data of such a diverse nature.

The basic assumptions of the model development of this paper are, firstly, the industries which are being considered can be aggregated in a transaction table. For the particular case under consideration, the A, B, and C type aggregation would be assumed to exist. Secondly, each industry has a time constant according to its ability to respond in kind to a particular production demand. Although it is recognized that some dispersion may exist within any particular aggregation, it is assumed however that one particular value can be assigned as a result of the total consideration for the particular industry. Given an initial economic aggregation, should the dispersion be too great, it is always possible to disaggregate to the point where the dispersion is brought within acceptable limits.

TERMINOLOGY

The system variables are defined as follows for $i,j=1,\dots,n$:

x_i - rate of production of the i-th commodity;

a_{ij} - the rate at which the i-th commodity is used to produce one unit of the j-th commodity per unit time;

b_{ij} - the stock of commodity i associated with the increase of level of activity producing commodity j;

s_i - supply rate of commodity i available for external consumption;

d_i - external demand rate for commodity i;

f_{ij} - matrix element giving the dynamics of the process.

$\underline{x}$, $\underline{s}$ and $\underline{d}$ are nx1 matrices, and $\underline{A}$, $\underline{B}$, $\underline{F}$ are nxn matrices. Based on the x_i and s_i definitions it is now possible to define for the static case, a set of production equations

$$x_i = \sum_{j=1}^{n} a_{ij} x_j + s_i \tag{7}$$

where a_{ij} is the i,j element of a production matrix $\underline{A}$ which is obtained from

$$a_{ij} = \frac{u_{ij}}{x_j} \tag{8}$$

where x_j is the gross output of industry j. Alternatively, $\underline{A}$ is given by

$$\underline{A} = \underline{US}^{-1} \tag{9}$$

where $\underline{S} = \text{diag}\ \{x_1, x_2, \ldots x_n\}$.

The matrix equation for the elements of (7) can now be written as

$$\underline{x} = \underline{A}\ \underline{x} + \underline{s} \tag{10}$$

$$(\underline{I}-\underline{A})\ \underline{x} = \underline{s} \tag{11}$$

The matrix $(\underline{I}-\underline{A})$ is known as the Leontief matrix if the matrix $\underline{A}$ is viable [2]. The system of equations (11) is known as the static Leontief model.

As mentioned previously, the basic analysis will in most cases be performed on the aggregated system. The affect of aggregation on the original system $(\underline{V},\underline{D})$ in the static Leontief model is shown below in equation (12).

$$(\underline{I}-\underline{G}'\underline{VGS}^{-1})\underline{x}=\underline{s} \tag{12}$$

Returning to the numerical example, $(\underline{V},\underline{D})$, of (1) with the aggregation, $\underline{G}'$, of (6), the numerical result for the equations of (11) or (12) is

$$(\underline{I}-\underline{A})\underline{x} = \begin{bmatrix} 19/22 & -2/24 & 0 \\ -2/22 & 18/24 & 0 \\ 0 & 0 & 16/24 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \tag{13}$$

The vector $\underline{x}$ gives rates of production for each of the industries. It is important in any solution, $\underline{x}$, that all of the x_i elements are non-negative, i.e., no negative rate of production. A matrix, $(\underline{I}-\underline{A})$, for which $\underline{x} \geq \underline{0}$ implies $(\underline{I}-\underline{A})\ \underline{x} \geq 0$ is said to be viable. The Hawkins-Simon test [4] for a viable matrix is that the matrix $(\underline{I}-\underline{A})$ have leading principle minors which are non-negative. Thus $(\underline{I}-\underline{A})^{-1}$, if it exists, will also be viable, and

$$\underline{x} = (\underline{I}-\underline{A})^{-1}\underline{s} \tag{14}$$

where $\underline{s} \geq 0$, insures that $\underline{x}$ will have all non-negative elements, $\underline{x} \geq 0$. It is this equation which determines the minimum production capacity required for each sector given a required external supply, $\underline{s}$.

A DYNAMIC LEONTIEF MODEL

Two of the most important economic variables are supply, $\underline{s}$, and demand, $\underline{d}$. In many dynamic Leontief models, [2,6,10] not both of these are included although it is often stated that if demand increases then supply will increase to meet the demand. Models with $\underline{s}(t) = \underline{d}(t)$ are often said to be in "dynamic equilibrium" [10]. The proposed model will include both variables and the variable causing the increase (decrease) of production of commodity will be the difference between demand and supply, i.e. "excess demand" ("excess supply").

Many proposed dynamic Leontief models [2,10] are unstable, in fact absolutely exponentially unstable. To a systems engineer, such an unstable model, which is capable of explaining so little, is an unacceptable one. The feedback models proposed in this paper under many conditions are stable, in fact, asymptotically stable.

The Leontief model used is an open one,[10] that is, one in which labor is not included and one which supplies production to external demands without taking into account pricing and pricing policies. The fundamental approach to the model is largely intuitionist with no particular attempt at this time to defend all its shortcomings or to explain all its advantages.

We start with a simple premise. The purpose of production is to provide commodities to be consumed. Some of the consumption is internal, that is, some of commodity i is used to produce commodity j. Some of the "consumption" is taken up in providing stock (especially for expanding production). Some of the consumption is external, that is, providing commodities to satisfy demands which are external to the productive system. External demand is regarded as an external input to the productive system. This demand determines the supply of each commodity which should be made available externally through the internal productive system.

Given an equilibrium condition where supply meets demand, an increase in demand to a new value should eventually result in an increase in supply until a new equilibrium is established. Thus, the input-output block diagram is as shown in Figure 2. The characteristics of the forward path are to be selected such that in the equilibrium case, the static Leontief model is obtained. The feedback element H, is to be chosen so that under equilibrium, supply = demand.

Two models are developed which differ only in that Model I does not assume pure integration in the forward path and Model II does. These differing assumptions affect the value of the feedback elements, but there is no essential difference between the closed loop dynamics of the two models.

MODEL I

In Model I, it is assumed that if there is an actuating error which is a constant value, then ultimately, the total rate of production will equal the actuating error. With this assumption, the dynamics are determined by

$$\dot{\underline{x}}(t) = -\underline{Fx}(t) + \underline{Fe}(t) \tag{15}$$

If $\underline{F}$ is diagonal, this equation gives a simple exponential relationship for the change in production of each commodity,

$$\dot{x}_i = -f_i \, x_i + f_i \, e_i(t) \tag{16}$$

and if e_i is a unit step, $e_i(t) = 1$ for $t > 0$,

$$x_i(t) = x_i(0)e^{-f_i t} + (1-e^{-f_i t}) \tag{17}$$

so that eventually, as $t \to \infty$, $x_i(\infty)=1=e_i(\infty)$.

Notice that the elements of $\underline{F}$ could be determined by applying step inputs of demand and observing the response of the total production, $\underline{x}(t)$. However, under this experiment, ameliorating influences such as pricing policies would have to be taken into account. The equations for production are given by (18) under reasoning similar to that giving (11).

$$\underline{x}(t) = \underline{Ax}(t) + \underline{B\dot{x}}(t) + \underline{s}(t) \tag{18}$$

where $\underline{x}(t)$ is the total produced output, $\underline{Ax}(t)$ is that used in the production process, $\underline{B\dot{x}}(t)$ is that placed into stock and $\underline{s}(t)$ is that which is available for satisfying external demand. Rewriting these equations yields

$$\dot{\underline{x}} = -\underline{Fx} + \underline{Fe} \tag{19}$$

$$\underline{s} = (\underline{I}-\underline{A})\underline{x} - \underline{B\dot{x}} \tag{20}$$

which is in a form familiar to most control engineers. The actuating error signal, $\underline{e}(t)$, is obtained as

$$\underline{e}(t) = \underline{d}(t) - \underline{Ks}(t) \tag{21}$$

and $\underline{K}$ is a feedback matrix to be selected. A signal flow diagram of the system is given in Figure 3.

These equations can be rewritten as

$$\dot{\underline{x}} = -\underline{Fx} + \underline{Fe} \tag{22}$$

$$\underline{s}(t) = (\underline{I}-\underline{A} + \underline{BF})\underline{x} - \underline{BFe} \tag{23}$$

$$\underline{e}(t) = \underline{d}(t) - \underline{Ks} \tag{24}$$

where $\underline{K}$ is to selected such that under equilibrium conditions (subscripted e):

(I) the demand and the supply are the same $(\underline{s}_e = \underline{d}_e)$;

(II) the static Leontief system is obtained, i.e., $\underline{s}_e = (\underline{I}-\underline{A})\underline{x}_e$.

If an equilibrium, $\underline{x}_e$, is established, $\dot{\underline{x}} = \underline{0}$, yielding, $\underline{x}_e = \underline{e}_e$. Under this condition,

$$\underline{s}_e = (\underline{I}-\underline{A})\underline{x}_e = (\underline{I}-\underline{A})\underline{e}_e \tag{25}$$

which satisfies (II) and

$$\underline{s}_e = (\underline{I}-\underline{A})\underline{d}_e - (\underline{I}-\underline{A})\underline{Ks}_e \tag{26}$$

Solving for $\underline{d}_e$ in terms of $\underline{s}_e$ yields

$$(\underline{I}-\underline{A})^{-1}(\underline{I} + (\underline{I}-\underline{A})\ \underline{K})\underline{s}_e = \underline{d}_e \tag{27}$$

Requiring the satisfaction of (I) and assuming this must hold for all $\underline{s}_e$ yields

$$\underline{K} = -\underline{A}(\underline{I}-\underline{A})^{-1} = -(\underline{I}-\underline{A})^{-1}\underline{A}. \tag{28}$$

Substituting this result in the original equation yields

$$\dot{\underline{x}} = -\underline{Fx} + \underline{Fe} \tag{29}$$

$$\underline{s} = (\underline{I}-\underline{A} + \underline{BF})\underline{x} - \underline{BFe} \tag{30}$$

$$\underline{e} = \underline{d} + \underline{A}(\underline{I}-\underline{A})^{-1}\underline{s} \tag{31}$$

These equations may be manipulated to yield the closed loop equations:

$$\dot{\underline{x}} = -F[\underline{I}-\underline{A}(\underline{I}-\underline{A+BFA})^{-1}(\underline{I}-\underline{A+BF})]\underline{x} + \underline{F}[\underline{I}-\underline{A}(\underline{I}-\underline{A+BFA})^{-1}\underline{BF}]\underline{d} \tag{32}$$

$$\underline{s} = (\underline{I}-\underline{A})(\underline{I}-\underline{A+BFA})^{-1}(\underline{I}-\underline{A+BF})\underline{x} - (\underline{I}-\underline{A})(\underline{I}-\underline{A+BFA})^{-1}\underline{BFd} \tag{33}$$

Under some cases, those for which the increase in stock is negligible, $\underline{B}=\underline{0}$ and the equations reduce to

$$\dot{\underline{x}} = -F(\underline{I}-\underline{A})\underline{x} + \underline{Fd} \tag{34}$$

$$\underline{s} = (\underline{I}-\underline{A})x \tag{35}$$

Because the closed loop system matrix is $-\underline{F}(\underline{I}-\underline{A})$, it is quite clear that for many important cases such as when $\underline{F}$ is diagonal, the closed loop system is asymptotically stable and $(\underline{I}-\underline{A})$ is a viable matrix.

MODEL II

The signal flow diagram for Model II is given in Figure 4. The equations describing the system are

$$\dot{\underline{x}} = \underline{Fe} \tag{36}$$

$$\underline{s} = (\underline{I}-\underline{A})\underline{x} - \underline{B}\dot{\underline{x}} \tag{37}$$

$$\underline{e} = \underline{d} \quad - \underline{Ks} \tag{38}$$

To satisfy (I) and (II), in equilibrium, $\dot{\underline{x}}=0 \Rightarrow \underline{e}_e = \underline{0} \Rightarrow \underline{K}=\underline{I}$. The closed loop equations are given by

$$\dot{\underline{x}} = -\underline{F}(\underline{I}-\underline{BF})^{-1}(\underline{I}-\underline{A})\underline{x} + \underline{F}(\underline{I}-\underline{BF})^{-1}\underline{d} \tag{39}$$

$$\underline{s} = (\underline{I}-\underline{BF})^{-1}(\underline{I}-\underline{A})\underline{x}-\underline{BF}(\underline{I}-\underline{BF})^{-1}\underline{d} \tag{40}$$

Once again, under the assumption that $\underline{B} \equiv \underline{0}$, these equations reduce to

$$\dot{\underline{x}} = -\underline{F}(\underline{I}-\underline{A})\underline{x} + \underline{Fd}(t) \tag{41}$$

$$\underline{s} = (\underline{I}-\underline{A})\underline{x} \tag{42}$$

which are exactly the same form as the equations for this case with Model I.

COMPARISON TO OTHER DYNAMIC MODELS

If the present models are compared against similar dynamic models, the obvious similarities exist such as linearity, etc. However, as mentioned previously, there are two distinguishing differences; the first is inclusion of the external demand and the productive supply for external consumption and the fact that the closed loop model could be stable whereas others [2,10] are essentially unstable. In fact, using the terminology here, both Chakravarty [2] and Sargan [10] came up with a dynamic model given by

$$\dot{\underline{x}} = \underline{B}^{-1}(\underline{I}-\underline{A})\underline{x} - \underline{B}^{-1}\underline{s} \tag{43}$$

This model is completely exponentially unstable since both $\underline{B}$ and $\underline{I}-\underline{A}$ are viable and thus have positive eigenvalues.

APPLICATIONS OF SYSTEM THEORY

Having the specific mathematical form of Model I or Model II allows any of the usual linear systems engineering type of investigations to be conducted. These would include sensitivity studies (examples in the Appendix) control studies, model verification, stochastic inputs, prediction, filtering, parameter optimization. This also allows the obvious generalizations which would permit $\underline{A}$, $\underline{B}$, $\underline{F}$, and $\underline{K}$ to be time varying. Of course much is left to be done and must wait further developments in this area.

REFERENCES

1. Bellman, R., Introduction to Matrix Analysis, McGraw-Hill, 1960, p. 175.
2. Chakravarty, S., Capital and Development Planning, MIT Press, 1969.
3. Chen, K., and Hoelscher, H.E., Mickle, M.H., Shah, K., Vogt, W.G., "Industrial Systems Modeling - I," to be published.
4. Hawkins, D., and Simon, H.A., "Note: Some Conditions of Macro-Economic Stability," Econometrica, 17 (1949), 245.
5. Jorgenson, D.W., "A Dual Stability Theorem," Econometrica, 28 (1960), 892-899.
6. Leontief, W.W., The Structure of the American Economy 1919-1939, Oxford Press, 1951.
7. Manne, A.S., and Rudra, A., "A Consistency Model of India's Fourth Plan," Sankhya, 27B (1965), 57.
8. Morishima, M., "Prices, Interest and Profits in a Dynamic Leontief System," Econometrica, 26 (1958), 358-380.
9. Oka, K., "Experiments with a Dynamic Model of a Less Developed Country," Proceedings of the Ninth Allerton Conference on Circuit and System Theory, 1971.
10. Sargan, J.D., "The Instability of the Leontief Dynamic Model," Econometrica, 26 (1958) 381-392.
11. Solow, R.M., "Competitive Valuation in a Dynamic Input-Output System," Econometrica, 27 (1959), 30-53.

APPENDIX

Consider a situation which would occur when new technologies could be applied to current production processes with the investment of capital. The obvious question is how this might most efficiently be accomplished and (incomplete) answers can be obtained statically by determining the sensitivity of the rate of production to changes in the $\underline{A}$ matrix for a specific $\underline{d}$ and dynamically by determining the sensitivity of a performance index to changes in the $\underline{A}$ matrix for a specific $\underline{d}$. In either case, the equations considered are those of (34)-(35) although more general analyses could be conducted.

The Static Case

Letting $\Delta\underline{A}$ be a perturbation in the $\underline{A}$ matrix, $\Delta\underline{x}$ be the resulting perturbation in the rate of productivity, and $\underline{d}_e$ some steady demand, yields after manipulation

$$\Delta\underline{x} = [(\underline{I}-\underline{A})^{-1}\Delta\underline{A}]\underline{d}_e \tag{44}$$

For example, if $\underline{d}_e = [1,0,0]'$, then x_1 is most sensitive to a_{11} in the numerical example considered.

The Dynamic Case

If in equations (34)-(35), $\underline{d}(t)$ is assumed to be a step input and the error $\underline{e}(t) = \underline{d}(t) - \underline{s}(t)$, it would be desirable to reduce $\underline{e}(t)$ for all time. A performance index P is given by

$$P = \int_0^\infty \underline{e}'(t)\ \underline{Re}(t)\ dt \tag{45}$$

where $\underline{R}$ is a real symmetric positive definite "weighting matrix." Then the error, $\underline{e}(t)$, can be obtained by solving

$$\dot{\underline{e}} = (\underline{A}-\underline{I})\underline{Fe} = \underline{Be} \tag{46}$$

with initial condition given by

$$\underline{e}(0) = \underline{d}(0+) - \underline{s}(0) \tag{47}$$

It turns out [1] that P can be obtained from

$$P = \underline{e}'(0)\ \underline{Xe}(0) \tag{48}$$

where $\underline{X}$ is the solution to the matrix equation

$$\underline{B}'\underline{X} + \underline{XB} = -\underline{R} \tag{49}$$

The sensitivity of P, ΔP, to changes in $\underline{B} = (\underline{A}-\underline{I})\underline{F}$ is given by

$$\Delta P = \underline{e}'(0)(\Delta\underline{X})\underline{e}(0) \tag{50}$$

where $\underline{X}$ is determined form (49) and $\Delta\underline{X}$ is determined as the solution to

$$(\underline{B}+\Delta\underline{B})'(\underline{X}+\Delta\underline{X}) + (\underline{X}+\Delta\underline{X})(\underline{B}+\Delta\underline{B}) = -\underline{R} \tag{51}$$

For example, if $\underline{R} = \text{diag}\ \{10, 1, .1\}$, $\underline{d} = [1, 1, 1]'$ for $t > 0$, $\underline{s}(0) = \underline{0}$, P turns out to be the most sensitive to changes in a_{11}.

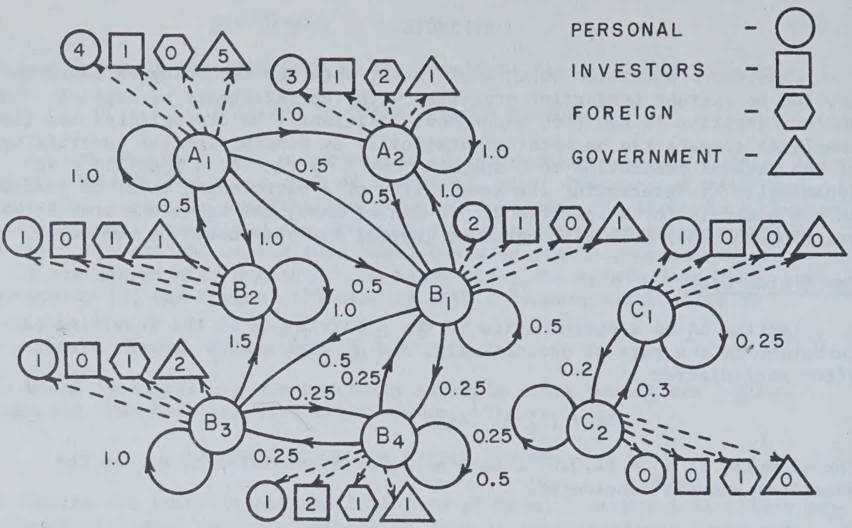

PERSONAL -
INVESTORS -
FOREIGN -
GOVERNMENT -
A1
A2
B1
B2
B3
B4
C1
C2
4
1
0
5
3
1
2
1
1.0
0.5
1.0
0.5
1.0
0.5
1.0
1.0
2
1
0
1
1
0
1
1
1
0
0
0
0.5
1.0
0.5
0.5
0.25
1.5
1.0
0.5
0.25
0.25
0.2
0.3
1
0
1
2
0.25
0.25
1.0
0.25
0.5
0
0
1
0
1
2
1
1

Figure 1. Example Industrial Complex.

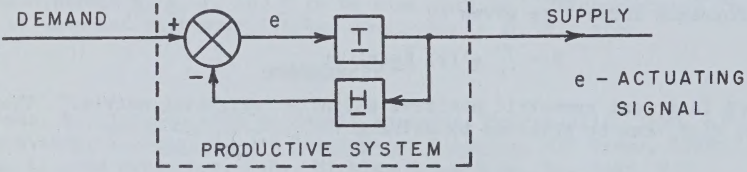

DEMAND
+
-
e
T
H
SUPPLY
PRODUCTIVE SYSTEM
e - ACTUATING SIGNAL

Figure 2. Input-Output Model.

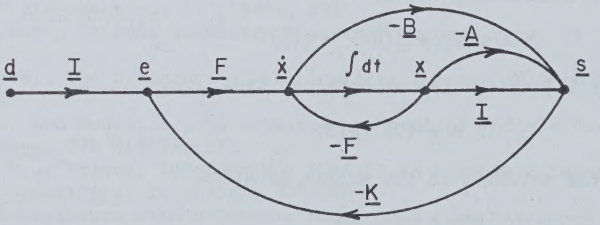

d
I
e
F
ẋ
∫dt
x
s
-B
-A
I
-F
-K

Figure 3. Model I.

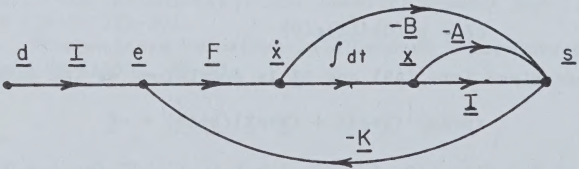

d
I
e
F
ẋ
∫dt
x
s
-B
-A
I
-K

Figure 4. Model II.

SUFFICIENT CONDITIONS FOR ALMOST CERTAIN STABILITY AND ASYMPTOTIC STABILITY OF A CLASS OF REGENERATIVE POPULATION SYSTEMS

WILLIAM D. O'NEILL
University of Illinois
Chicago, Illinois

ABSTRACT

Data from such diverse populations as the human population of the United States and the redwood population of northern California suggests that the dynamics of these populations can be captured by the same mathematical model.[1,2] Fitting data in a mean square sense results in a model dependent on three parameters: a rate constant, a feedback gain and the dimension of the system state vector. For this model there are derived:

1. Necessary and sufficient conditions on the parameters (considered to be deterministic) for asymptotic and demographic stability of the null state.

2. Sufficient conditions on the parameters (considered to be random) for almost certain asymptotic and demographic stability of the null state.

THE MODEL

Let m_k be the number of births per unit time per adult entity of age k. Let l_i be the probability of survival at time i of an entity born at time 0. Let r_k be the birth rate (entities per unit time) at time k. If h_k is a function depending on conditions for $k<0$ then

$$r_k = h_k + \sum_{j=0}^{k-1} m_{k-j} l_{k-j} r_j \qquad (1)$$

(1) defines a stationary, linear, deterministic, discrete-time, feedback system with weighting sequence $m_u l_u \overset{d}{=} g_u$.

Data for the human population of the United States and the redwood population of northern California was used to derive a representative form for g_u. It was assumed that

$$Z\{g_u\} = \frac{G(1-a)^2}{z^T(z-a)^2} . \qquad (2)$$

T was fixed while G and a were varied to minimize the mean square error between model and data. The results are shown in figure 1. Let x_k = col(r_k, r_{k+1}, r_{k-1}, . . . r_{k-T}) then (1) becomes

$$x_{k+1} = Cx_k \qquad (3)$$

with C a(T+2) x (T+2) matrix given by

$$C = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ -a^2 & 2a & 0 & \cdot & \cdot & \cdot & G(1-a)^2 \\ 1 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdot & 0 \\ & & & \vdots & & & \\ 0 & 0 & \cdot & \cdot & \cdot & 1 & 0 \end{bmatrix}$$

DETERMINISTIC STABILITY

The characteristic equation of C is

$$f(z) = z^n(z-a)^2 - G(1-a)^2 = 0 \qquad (4)$$

where $n \overset{d}{=} T$. Constraints on n, a and G arise from the nature of population dynamics. Disallowing negatives entities requires $G \geq 0$, $a \geq 0$. The discrete nature of the system implies $n \geq 0$ and integer. If entities are to have finite life spans or finite reproductive spans then $a < 1$. With these assumptions we have

Theorem 1

The necessary and sufficient condition that C have all of its eigenvalues less than unity in absolute value is $G < 1$.

Proof: If all the eigenvalues of C satisfy $|z| < 1$ then the Schur-Cohn theorem[3] requires $f(1) > 0$. From (4) $f(1) = (1-a)^2 - G(1-a)^2$. Since $a \neq 1$, $G < 1$.

Suppose $G < 1$ then by (4) $|z|^n < \dfrac{(1-a)^2}{|z|^2 + a^2 - 2a\mathrm{Re}z}$ which implies $|z|^n(|z|^2 + a^2 - 2a\mathrm{Re}z) < (1-a)^2$. But $(|z|-a)^2 \leq |z|^2 + a^2 - 2a\mathrm{Re}z$ so that $|z|^n(|z|-a)^2 < (1-a)^2$. Assume $|z| > 1$ then $|z|^n > 1$ and $(|z|-a)^2 > (1-a)^2$ therefore $|z|^n(|z|-a)^2 > (1-a)^2$. This contradiction establishes $G < 1$ implies $|z| < 1$.

Theorem 1 is concerned with the stability of the null solution of (3). Asymptotic stability of $x = 0$ implies extinction of the species and therefore a question of more practical interest is the demographic stability of (3). Specifically, we seek conditions under which the null state of (3) is stable but not asymptotically stable.

Theorem 2

The necessary and sufficient condition that the null state of (3) be stable but not asymptotically so is $G = 1$.

Proof: Assume the null state is stable. Therefore

A. C has one and only one unity eigenvalue and n-1 distinct complex eigenvalues whose moduli are less than one or

B. Some of the eigenvalues of C are distinct, complex and unity in moduli. Since (4) implies $G = \frac{|z|^n(|z|^2+a^2-2aRez)}{(1-a)^2}$ case A requires G = 1.

Case B requires $G = \frac{(1+a^2-2aRez)}{(1-a)^2}$. But $Rez<1$ implying $1+a^2-2aRez>(1-a)^2$ which requires $G>1$. By theorem 1 this is a contradiction. Thus case A must prevail so that stability of the null solution implies G = 1.

Suppose G = 1 then by (4) $f(1) = 0$. Therefore z=1 is an eigenvalue of C. Replacing the strict inequalities with inequalities in the proof of theorem 1 results in the same contradiction so that G = 1 implies $|z|\leq 1$. Dividing (4) by z-1 and putting z = 1 in the resulting polynomial results in the constraint $a=1+n/2$. By assumption $0\leq a<1$, $n\geq 0$, so that z = 1 is not a repeated eigenvalue.

STOCHASTIC STABILITY

Of the three parameters selected to characterize population dynamics the feedback gain, G, is the most likely to be non-stationary. This is because G is the net reproductivity ratio of the population and, for populations capable of decision making, G depends on many exogenous influences. We therefore partition the matrix C of (3) so that (3) becomes

$$x_{k+1} = (A+B_k)x_k \tag{5}$$

where the only non-zero element of B_k is $G_k(1-a)^2$. The A matrix has the characteristic equation

$$g(z) = z^n(z-a)^2 = 0 \tag{6}$$

so that it is asymptotically stable. We further assume the non-zero element of B_k is measurable and continuous with probability one, strictly stationary, ergodic and $E(|B_k|)$ exists. Theorem 3 will require the lemma

Lemma 1

If u_k, $v_k\geq 0$, b a positive constant and $u_k\leq b+\sum_{j=0}^{k-1} u_j v_j$ for all k then $u_k\leq b\prod_{j=1}^{k-1}(1+v_j) \leq b(1+\frac{1}{k}\sum_{j=0}^{k-1} v_j)^k$

Proof: By assumption $u_{k+1}\leq b+\sum_{j=0}^{k} u_j v_j = u_k(1+v_k)$. Also $u_o\leq b$.

Thus

$$u_1 \leq u_0(1+v_0)$$

$$u_2 \leq u_1(1+v_1) \leq u_0(1+v_0)\ (1+v_1)$$

$$\cdot$$

$$\cdot$$

$$\cdot \quad u_k \leq u_0 \prod_{j=0}^{k-1} (1+v_j) \leq b \prod_{j=0}^{k-1} (1+v_j)$$

By the geometric inequality

$$u_k \leq b \prod_{j=0}^{u-1} (1+v_j) \leq b\Big(1+\frac{1}{k}\sum_{j=0}^{k-1} v_j\Big)^k.$$

Lemma 1 can be applied to the solution of (5) since

$$|x_k| \leq |A^k x_0| + \sum_{j=0}^{k-1} |A^{k-j-1}| \cdot |B_j| \cdot |x_j|. \tag{7}$$

A is asymptotically stable so there exists a number $0<c<1$ and a $h>0$, depending on A, such that $|A^k|<c^k h$

From (7)

$$c^{-k}|x_k| \leq \Big(|x_0| + \sum_{j=0}^{k-1} \frac{|B_j|}{c}\, c^{-j}|x_j|\Big) h \tag{8}$$

In lemma 1 we identify

$$(u_k, v_k, b) = \Big(c^{-k}|x_k|, h\frac{|B_k|}{c}, |x_0| h\Big)$$

so that

$$c^{-k}|x_k| \leq |x_0| \Big(1+\frac{1}{k}\sum_{j=0}^{k-1} h\frac{|B_j|}{c}\Big)^k h \tag{9}$$

Finally

$$|x_k| \leq |x_0| \Big(c + \frac{1}{k}\sum_{j=0}^{k-1} h|B_j|\Big)^k h \tag{10}$$

Clearly $|x_k| \to 0$ as $k\to\infty$ if $\Big(c+\frac{1}{k}\sum_{j=0} h|B_j|\Big) < 1$.

Also $\lim_{k\to\infty} \frac{1}{k}\sum_{j=0}^{k-1} |B_j| = E(|B_k|)$ with probability one and $|B_k| = |G_k|\ (1-a)^2$.

Thus there exists a $d>0$ such that $|x_k| \to 0$ with probability one if $E(|G_k|) < \dfrac{1-c}{h(1-a)^2} = d$ which proves

Theorem 3

Let A in (5) be asymptotically stable and suppose $E(|G_k|)$ exists. Then there exists a constant $d(A) > 0$ such that $E(|G_k|) < d$ implies the null solution of (5) is almost certainly asymptotically stable.

Any population satisfying theorem 3 is doomed to extinction. Survival possibilities in a non-zero and bounded population (a demographically stable condition) are found in the conditions of theorem 2. However the stability of theorem 2 appears to be marginal in that the solution may not be able to tolerate disturbances. To show this let $G_k = 1+R_k$ with R_k satisfying the same conditions imposed on B_k just previous to lemma 1. A in (5) is now C of (3) with G = 1 and the only non-zero element of B_k in (5) is just R_k. A now satisfies theorem 2 so there exists a constant $M>0$ such that $|A^k|<M$.

Thus

$$|x_k| \leq M|x_o| + \sum_{j=0}^{k-1} M|R_j|\cdot|x_j|. \qquad (11)$$

Applying lemma 1 to (11)

$$|x_k| \leq M|x_o| \left(1 + \frac{M}{k} \sum_{j=0}^{k-1} |R_j|\right)^k. \qquad (12)$$

For non-asymptotic stability we require $1+ME(|R_k|) = 1$ or $E(|R_k|) = 0$. But R_k was assumed stationary so that $E(|R_k|) = 0$ implies $|R_k| \equiv 0$ for all k with probability one.

COMMENTS

Applying the above deterministic results to the human and redwood populations we note that both systems have unstable equilibria. The question naturally arises as to why the earth is not saturated with people and redwood trees. The answer is because the model is only of the system responsible for producing new entities. The total population, P_k, at any specific time k is given by

$$P_k = \sum_{i=0}^{\infty} l_i r_{i-k}.$$

The problem of people saturation is a real one since l_i for people ranges between 0.97 for i = 0 and decreases monotonically to zero for increasing i[4]. Additionally people have no significant predators to keep an unstable birth output in check. For redwoods however the maximum l_i is about 0.01 and these trees have many natural predators (including man) at work to check their population. An interesting aspect of the results that deserves more attention is the knife-edge stability (G=1) of a stable solution itself. This solution, tottering between extinction and chaos, could possibly be made asymptotically stable by feeding back instantaneous and past birth rates. The feedback information would be used to make G a function of the system state and thus make the overall system nonlinear. Hopefully the resultant system would have a non-zero, asymptotically stable equilibrium whose magnitude could be adjusted by varying the amount of feedback.

ACKNOWLEDGEMENT

The author gratefully acknowledges the assistance of Dr. Joseph Engel and Mr. Nicholas Jermihov in reaching some of the conclusions of this paper.

BIBLIOGRAPHY

1. N. Keyfitz, Introduction to the Mathematics of Populations, Addison Wesley, 1968.
2. C.A. Bosch, "Redwoods: A Population Model", Science, Vol. 172, pp. 345-349, April 23, 1971.
3. E.I. Jury, "A Stability Test for Linear Discrete Systems Using a Simple Division", Proc. IRE, Vol. 49, No. 12, p. 1948, December, 1961.
4. W.D. O'Neill, "A Systems Model of the People Generative Process". Proc. 5th Annual Princeton Conf. on Info. and Syst., March, 1971.

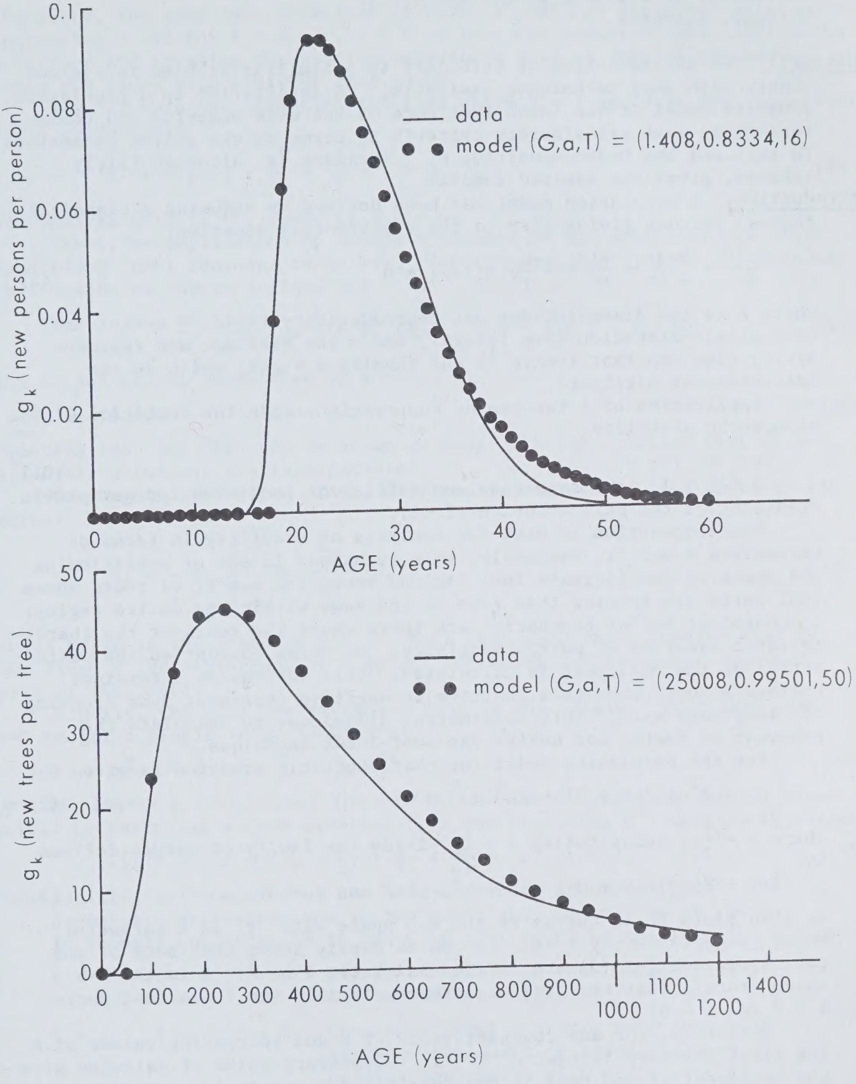

0.1
0.08
0.06
0.04
0.02
g$_k$ (new persons per person)
data
model (G,a,T) = (1.408,0.8334,16)
0
10
20
30
40
50
60
AGE (years)
50
40
30
20
10
0
g$_k$ (new trees per tree)
data
model (G,a,T) = (25008,0.99501,50)
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1400
AGE (years)

Figure 1.

AN APPLICATION OF D-DECOMPOSITION TO THE STABILITY ANALYSIS OF A POPULATION MODEL

NICHOLAS JERMIHOV
University of Illinois at Chicago Circle
Chicago, Illinois

Abstract: The determination of stability is a familiar problem in systems theory with many techniques available. In application to a population dynamics model it was found that none of the more powerful and common techniques gave satisfactory criteria in terms of the system parameters To this end the D-decomposition of Y. Neimark,[2], although fairly unknown, gives the desired results.

Introduction: A population model has been derived by assuming a classical renewal process giving rise to the differential equation:

$$\frac{d^2 r}{dt} + \frac{2}{\tau}\frac{dr}{dt} + \frac{1}{\tau^2} r - \frac{K}{\tau^2} x(t-\theta) = 0 \qquad (1)$$

Where K is the dimensionless net reproductivity ratio, θ social and biological maturation time (years), and τ the marriage and reproductivity time constant (years^{-1}) and finally $r = r(t)$ which is the instantaneous birthrate.

Results: Application of a theorem by Pontryagin yields the condition C1 for asymptotic stability.

$$\theta > 0, \ \tau > 0 \qquad (C1)$$

$0 \leq K < 1$ is a necessary and sufficient condition for asymptotic stability of the null solution of (1).

D-decomposition is used for analysis of stability in terms of parameters θ and τ. Basically, this technique is one of partitioning the space of coefficients into regions where the number of roots whose real parts are greater than zero is the same within the entire region. A convenient set of boundaries are those where the roots of the characteristic equation of purely imaginary. On these boundaries the differential of the real part is calculated. This derivative determines whether a root receives a positive or negative decrement upon crossing the imaginary axis. This information allows one to calculate the behavior of roots, not unlike the root-locus technique.

For the population model the characteristic equation is given by:

$$Z^2 + 2bZ + b^2 - Kb^2 e^{-Z} = 0 \qquad (2)$$

where $b = \frac{\theta}{\tau}$. Substituting $Z = iy$ yields the family of curves defined by:

$$2by + Kb \operatorname{Sin} y = 0 \qquad -y^2 + b^2 - Kb^2 \cos y = 0. \qquad (3)$$

We also plots these curves in the K-b space with 'y' as a parameter whose range is $-\infty < y < +\infty$. It can be easily shown that none of the curves of equation (3) are located below the line $K = 1$ in the K-b space (recall that the area of interest is only the I quadrant where $K \geq 0$ and $b \geq 0$).

Therefore, for any constant value of b and increasing values of K the first boundary is that of $K = 1$. Elementary rules of calculus give the increment of the real of the root for the system $Z = x + iy$ $f(Z) = Z^2 + 2bZ + b^2 - Kb^2 e^{-Z}$ as:

$$dx = -\operatorname{Re} \frac{\partial f / \partial K}{\partial f / \partial Z} \, dK \qquad (5)$$

substituting these partial derivatives and evaluating at the boundary we arrive at:

$$dx = \frac{b}{2+b}\, dK. \tag{6}$$

Therefore, the real part of a root receives a positive increment with increasing K and for $K = 1 + \varepsilon$, $\varepsilon > 0$ we have one unstable root. Increasing K further and crossing the other boundaries a pair of complex conjugate roots are added with positive real parts.

Now, we turn our attention the case where K = 1 yielding the characteristic polynomial:

$$Z^2 + 2\frac{\theta}{\tau} Z + \frac{\theta^2}{\tau^2} - \frac{\theta^2}{\tau^2} e^{-Z} = 0 \tag{7}$$

and analyze this in the space of θ and τ.

First, by application of Rouche's theorem we can show that the right hand plane (RHP) contains no zeroes. Substituting $Z = iy$ into (7) results in a family of curves defined by:

$$-y^2 + \frac{\theta^2}{\tau^2} - \frac{\theta^2}{\tau^2} \cos y = 0 \quad (A) \qquad 2 + \frac{\theta}{\tau} \sin y = 0 \quad (B) \tag{8}$$

and an additional curve that of y = 0.

$$\theta = 2\tau. \tag{9}$$

Equation (8A) and (8B) can be shown to have only one solution that of y=0, all other solutions are incompatible!

Once again evaluation of the differential increment in x where $Z=x+iy$ gives:

$$dx = -\operatorname{Re} \frac{2\frac{\theta}{\tau^2} - 2\frac{\theta}{\tau^2}}{2\frac{\theta}{\tau} \qquad \frac{2\theta}{\tau^2}}\, d\theta \tag{10}$$

which holds on the boundary $\theta = 2\tau$. Note that dx=0 which implies that for any fixed τ the variations of θ do not increment nor decrement the real part of the roots.

Finally, the case Z=x reveals a root at the origin. Note that where y=0 we get a single root, looking at:

$$x^2 + 2\frac{\theta}{\tau} x + \frac{\theta^2}{\tau^2} - \frac{\theta^2}{\tau^2} e^{-x} = 0 \tag{11}$$

x=0 is clearly a root. That the multiplicity is of order one can be ascertained by factoring x from equation (11) and rewriting e^{-x} as a power series:

$$x[X + 2\frac{\theta}{\tau} + \frac{\theta^2}{\tau^2}(x - \frac{x^2}{2!} + \frac{x^3}{3!} - \frac{x^4}{4!} + \ldots..) = 0 \tag{12}$$

The resulting polynomial does not have any roots at the origin.

Asymptotic approximations to the roots show no multies of roots.

In summary condition C2 is added to that of C1 to complete the analysis.

K = 1	$\theta > 0$, $\tau > 0$	Stable (no roots on the imaginary axis and one at the origin)	
K > 1	$\theta > 0$ $\tau > 0$	Unstable, first a single real root with positive real part is added followed by complex conjugates with positive real parts as K is increased.	(C2)

Reference:

1. "A Systems Model of the People Generative Process," Proceedings of the Fifth Annual Princeton conference on Information Sciences and Systems, March, 1971, (in press).
2. "The Structure of the D-decomposition of the Space of Quasi-polynomials," Doklady Akodemii Nauk SSSR, 1948, (in Russian).

AN INTERACTIVE FACILITY FOR EXAMINING DEVELOPMENT POLICIES OF LOCAL GOVERNMENT

E. J. Cristiani, R. E. Goldman, Santa Clara County
(California) Planning Department
R. J. Evey, P. E. Mantey, IBM Research Division, San Jose,
California

Abstract

This paper describes an interactive system incorporating a model of the growth of Santa Clara County, California. The system provides hands-on control (via an IBM 2250 display) of input, output, changes in key data, running of the model, and various displays. The model used in this way can aid a decision maker in evaluating policies which influence the pattern of growth.

Summary

Land development in many areas, including Santa Clara County, California, has been left largely to the private sector, usually accommodated, if not abetted, by local governments. Out of concern for the costs of providing urban services and preserving urban amenities many cities (in particular, San Jose) are becoming interested in developing ways to influence more directly the pattern of land development.

This paper describes the results to date of a joint study among IBM, the County of Santa Clara, and the City of San Jose. The study is intended to result in an instrument which can be continuously refined for the evaluation of the effectiveness of governmental control policies on land development. Toward this objective, three related areas are being studied:

1. Models for projecting future development, especially for relating local governmental policies to the pattern of development;
2. Data base availability and requirements for use in decision making regarding policies and public improvements; and
3. Display and data retrieval to make the model output and data base useful to decision makers.

A general diagram of the context of this effort is given in Figure 1. This diagram is oversimplified since in our case the driving function is really provided by three separate programs as shown in Figure 2. The data base characterizes the base-year state including physical conditions of land, existing facilities, and present state of development, as well as present state of economic conditions in the areas under study. The controls are those governmental policies which provide various (positive or negative) incentives to development. Simulation forecasts the state of future development, based on the inputs indicated. The display function allows user appraisal of the effects of various control policy choices, presenting an interpretation of development at selected future times. The policy innovator evaluates the resulting pattern of land development, and if desired, reruns the system utilizing different controls. In this way, by a process primarily of trial and error, a set of controls which affect development can be examined.

A set of models projecting land use by zone (the county is divided into 334 zones based on census tracts) for residential and commercial purposes is now running on an IBM model 360/195 (Figure 2). These models were developed by the Planning Department of Santa Clara County as part of a transportation study and have been revised by IBM. Calibration and sensitivity studies are underway as well as plans for major extensions and improvements.

A display program has been developed based on a zonal map of the county.

This program permits the policy innovator to ask questions via the 2250 keyboard relating the values of the variables in any data file. The program interprets these statements, searches the prescribed file to determine which zones satisfy these statements, and then displays these zones on the 2250 screen. Several series of inclusive questions can be asked at once. A different alphanumeric symbol, chosen by the user, will distinguish the several sets of zones on the screen. Another option is the elimination of common boundaries on the screen of adjacent zones which satisfy the same statements. Additional options allow display and alteration of numerical data by zone or by variable.

These programs are controlled by the user from the 2250. The user can change data relating to policy decisions (such as land availability or residential density) via the keyboard and light pen of the 2250, rerun the models, and use the display for examining the resulting pattern of the development. The model projects development in five-year steps. At the discretion of the user, the results of any step can be stored in the data base. Each of the files, thus, created is then available for display or for input as base year conditions for another iteration of the model. In this way, alternative growth patterns can be developed and compared by the user.

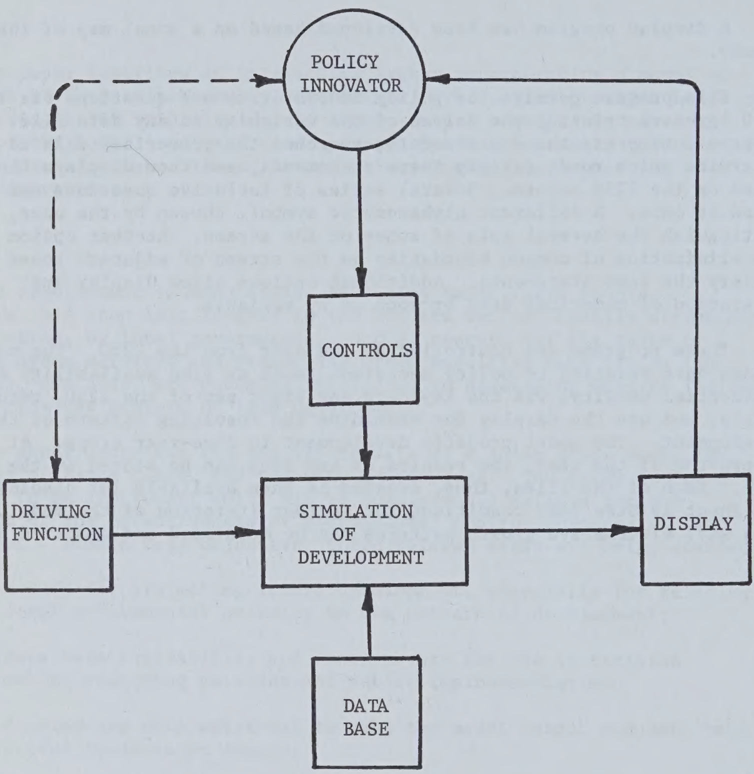

POLICY INNOVATOR
CONTROLS
DRIVING FUNCTION
SIMULATION OF DEVELOPMENT
DISPLAY
DATA BASE

Figure 1

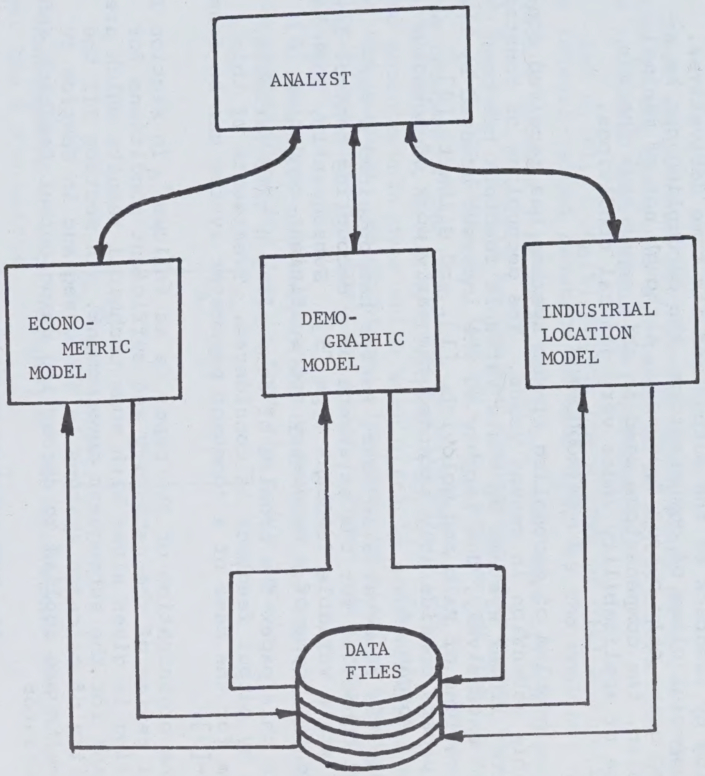

ANALYST
ECONO-METRIC MODEL
DEMO-GRAPHIC MODEL
INDUSTRIAL LOCATION MODEL
DATA FILES

Figure 2

DECOUPLING IN LINEAR TIME-VARIABLE SYSTEMS USING OUTPUT FEEDBACK

SAHJENDRA N. SINGH and WILSON J. RUGH
Department of Electrical Engineering
The Johns Hopkins University
Baltimore, Maryland 21218

ABSTRACT

For a linear time-variable plant which can be decoupled by state variable feedback, the problem of decoupling by output feedback is discussed. The problem is solved by considering the decoupling problem for a system composed of a compensator in cascade with the plant. The use of the compensator allows decoupling by feedback of the output and its time derivatives. For a special class of compensators, the decoupling can be achieved using only output feedback. Although not of minimal dimension, the compensators used in this paper have the advantage of applicability under very general conditions.

I INTRODUCTION

The problem of decoupling linear systems has received considerable attention in recent years. The decoupling of constant parameter linear systems by state variable feedback has been studied extensively, due largely to the interest aroused by the contributions of Falb and Wolovich, [1], and Gilbert, [2]. These papers considerably extended the early work of Rekasius [3], and Morgan [4].

For the time-variable case, Porter has obtained a sufficient condition for the existence of a decoupling control law of the state variable feedback type [5]. Subsequently, Rugh has derived two forms of a necessary and sufficient condition [6].

In this paper the problem of decoupling a time-variable system by output feedback is considered. Treatments of this problem for the case of a constant parameter system can be found in [7]-[9].

The organization of the paper is as follows. In Section II a brief review of the necessary and sufficient conditions for decoupling is given along with some technical results which are important for the subsequent development. In Section III the decoupling of cascaded systems is discussed and in Section IV the results are applied to decoupling using output feedback and a compensator.

II PRELIMINARY DEVELOPMENT

We shall consider the linear time-variable plant

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) \end{aligned} \tag{1}$$

which is defined for $t\in[t_0,t_1]$. The state vector $x(t)$ is $n \times 1$ and the input and output vectors, $u(t)$ and $y(t)$ respectively, are $m \times 1$. For simplicity we assume that the elements of A, B, and C are real-valued functions of time which are differentiable any finite number of times on $[t_0,t_1]$. (The precise number of continuous derivatives required for the various results of the paper can be easily determined.)

The control laws of interest are those of the form

$$u(t) = F(t)x(t) + G(t)w(t) \tag{2}$$

where $w(t)$ is an $m \times 1$ input vector and where $G(t)$ is nonsingular at each $t \in [t_0, t_1]$. We wish to choose a control law (2) which decouples the plant (1). That is, which gives a closed-loop system with a diagonal impulse response matrix [6].

In considering this problem, the following two operators play a crucial role:

$$\begin{aligned} L_A(\cdot) &= \frac{d}{dt}(\cdot) + (\cdot)A(t) \\ R_A(\cdot) &= \frac{d}{dt}(\cdot) - A(t)(\cdot). \end{aligned} \tag{3}$$

In these expressions $A(t)$ is the given plant matrix and the arguments of the operators are taken to be any matrices such that the indicated expressions are conformable.

Let $c_1(t), \ldots, c_m(t)$ denote the m rows of $C(t)$ and define the integer-valued function of time $d_i(t)$ as the smallest non-negative integer j at each $t \in [t_0, t_1]$ such that $L_A^j(c_i)(t)B(t) \neq 0$, $i=1,2,\ldots,m$. It will be assumed throughout this paper that each $d_i(t)$ is a constant, d_i, for $t \in [t_0, t_1]$ and we let $d^* = \min_{i=1,\ldots,m}(d_i)$. (The results of the paper can be extended to the case of nonconstant $d_i(t)$ by dividing $[t_0, t_1]$ into subintervals over which each $d_i(t)$ is constant and considering the various subproblems. However, in piecing together the subproblems some difficulties arise. These are discussed in Section V.)

A necessary and sufficient condition for the existence of a decoupling control law is that the $m \times m$ matrix

$$B^*(t) = \begin{bmatrix} L_A^{d_1}(c_1)(t)B(t) \\ \cdot \\ \cdot \\ \cdot \\ L_A^{d_m}(c_m)(t)B(t) \end{bmatrix} \tag{4}$$

be nonsingular at each $t \in [t_0, t_1]$. An equivalent condition is that the $m \times m$ matrix

$$D^*(t) = \begin{bmatrix} c_1(t)R_A^{d_1}(B)(t) \\ \cdot \\ \cdot \\ \cdot \\ c_m(t)R_A^{d_m}(B)(t) \end{bmatrix} \tag{5}$$

be nonsingular at each $t \in [t_0, t_1]$.

The following relationships involving the operators L_A and R_A will be useful in the sequel. (The proofs are discussed in the Appendix.)

Lemma 1 For all nonnegative integers j and k,

$$L_A^j(c_i)(t)R_A^k(B)(t) = (-1)^k L_A^{j+k}(c_i)(t)R_A^0(B)(t) + \sum_{q=1}^{k} (-1)^{k+q} \frac{d}{dt}\left[L_A^{j+k-q}(c_i)(t)R_A^{q-1}(B)(t)\right], \qquad (6)$$

where, as customary, $L_A^0(c_i)(t) = c_i(t)$ and $R_A^0(B)(t) = B(t)$.

Corollary 1 For the nonnegative integers j and k,

$$L_A^j(c_i)(t)R_A^k(B)(t) = \begin{cases} 0, & j+k < d_i \\ (-1)^k L_A^{d_i}(c_i)(t)B(t), & j+k = d_i \end{cases} \qquad (7)$$

Lemma 2 Suppose $B^*(t)$ is nonsingular at each $t \in [t_0,t_1]$. Then $m + \sum_{i=1}^{m} d_i \le n$ and, furthermore, if equality holds then (A,C) is uniformly observable on $[t_0,t_1]$. [10]

Corollary 2 Suppose $B^*(t)$ is nonsingular at each $t \in [t_0,t_1]$ and $d_1 = d_2 = \ldots = d_m = d^*$. Then $d^* \le (n-m)/m$. Also, if (A,C) is uniformly observable on $[t_0,t_1]$ with observability index ν_0, then $d^* \le \nu_0 - 1$.

Results involving uniform controllability and the controllability index ν_c which are similar to those above can be derived. However, since these results are not needed in the sequel they are omitted.

III DECOUPLING OF CASCADED SYSTEMS

We shall consider compensators which are m-input, m-output systems of order n_c and are described on $[t_0,t_1]$ by

$$\begin{aligned} \dot{z}(t) &= D(t)z(t) + E(t)r(t) \\ v(t) &= H(t)z(t). \end{aligned} \qquad (8)$$

This system is assumed to have the same differentiability properties as the plant (1). We define the compensator parameters $k_i(t)$ at each $t \in [t_0,t_1]$ as the smallest nonnegative integer j such that $L_D^j(h_i)(t)E(t) \neq 0$, where $h_i(t)$ is the i^{th} row of H(t), i = 1,...,m. We assume that each $k_i(t)$ is a constant k_i for $t \in [t_0,t_1]$ and we set $k^* = \min_{i=1,\ldots,m} (k_i)$.

Suppose that the compensator (8) is placed in cascade with the plant (1) so that u(t) = v(t). The resulting system can be described as

$$\begin{aligned}\dot{\tilde{x}}(t) &= \tilde{A}(t)\tilde{x}(t) + \tilde{B}(t)r(t)\\ y(t) &= \tilde{C}(t)\tilde{x}(t),\end{aligned} \tag{9}$$

where, in partitioned form,

$$\tilde{x}(t) = \begin{bmatrix} x(t) \\ z(t) \end{bmatrix}, \quad \tilde{A}(t) = \begin{bmatrix} A(t) & B(t)H(t) \\ 0 & D(t) \end{bmatrix}$$
$$\tilde{B}(t) = \begin{bmatrix} 0 \\ E(t) \end{bmatrix}, \quad \tilde{C}(t) = \begin{bmatrix} C(t) & 0 \end{bmatrix} \tag{10}$$

In order to consider the decoupling of the composite system (9), we need the following results which relate $L_{\tilde{A}}(\tilde{c}_i)(t)$, where $\tilde{c}_i(t)$ is the i^{th} row of $\tilde{C}(t)$, to the operators L_A and L_D, where $L_A(\cdot)$ is defined in (3) and $L_D(\cdot) = \frac{d}{dt}(\cdot) + (\cdot)D(t)$.

Lemma 3 For any nonnegative integer j,

$$L^j_{\tilde{A}}(\tilde{c}_i)(t) = \left[L^j_A(c_i)(t) \vdots \sum_{q=0}^{j-1} L_D^{j-q-1}(L_A^q(c_i)(t)B(t)H(t))(t) \right] \tag{11}$$

Corollary 3 For any nonnegative integer j,

$$L^j_{\tilde{A}}(\tilde{c}_i)(t)\tilde{B}(t) = \begin{cases} 0\ , & j < d_i + 1 \\ \displaystyle\sum_{q=d_i}^{j-1} L_D^{j-1-q}(L_A^q(c_i)(t)B(t)H(t))(t)E(t), & j \geq d_i + 1 \end{cases} \tag{12}$$

Lemma 4 If M(t) and N(t) are matrices with elements that are differentiable any finite number of times, and if the dimensions of M(t) and N(t) are such that the indicated operations are conformable, then for any nonnegative integer j,

$$L_D^j(M(t)N(t))(t) = \sum_{q=0}^{j} \binom{j}{q} \left[\frac{d^q}{dt^q} M(t)\right] L_D^{j-q}(N)(t) \tag{13}$$

Corollary 4 For any nonnegative integers j and p,

$$L_D^j(L_A^p(c_i)(t)B(t)H(t))(t)E(t) = \begin{cases} 0\ , & j < k^* \\ L_A^p(c_i)(t)B(t)L_D^{k^*}(H)(t)E(t)\ , & j = k^* \end{cases} \tag{14}$$

We are interested in decoupling the composite system (9) using a control law of the form

$$\begin{aligned} r(t) &= \tilde{F}(t)\tilde{x}(t) + \tilde{G}(t)w(t) \\ &= F_1(t)x(t) + F_2(t)z(t) + \tilde{G}(t)w(t) \end{aligned} \tag{15}$$

where $\tilde{G}(t)$ is nonsingular at each $t \in [t_0, t_1]$. To this end, define $\tilde{d}_i(t)$ at each $t \in [t_0, t_1]$ as the smallest nonnegative integer j such that $L^j_{\tilde{A}}(\tilde{c}_i)(t)\tilde{B}(t) \neq 0$, $i = 1, \ldots, m$, and let

$$\tilde{B}^*(t) = \begin{bmatrix} L_{\tilde{A}}^{\tilde{d}_1(t)}(\tilde{c}_1)(t)\tilde{B}(t) \\ \cdot \\ \cdot \\ \cdot \\ L_{\tilde{A}}^{\tilde{d}_m(t)}(\tilde{c}_m)(t)\tilde{B}(t) \end{bmatrix} \tag{16}$$

Theorem 1 Suppose that the compensator (8) can be decoupled. That is, suppose that the m x m matrix

$$E^*(t) = \begin{bmatrix} L_D^{k_1}(h_1)(t)E(t) \\ \cdot \\ \cdot \\ \cdot \\ L_D^{k_m}(h_m)(t)E(t) \end{bmatrix} \tag{17}$$

is nonsingular at each $t\in[t_0,t_1]$. Also, suppose that $k_1=k_2= \ldots = k_m = k^*$. Then

$$\tilde{d}_i(t) = \tilde{d}_i = d_i + k^* + 1, \quad i = 1,\ldots,m \tag{18}$$

$$\tilde{B}^*(t) = B^*(t)E^*(t) \ , \quad t\in[t_0 t_1] \ . \tag{19}$$

Thus the composite system can be decoupled if and only if the plant (1) can be decoupled.

The compensator condition $k_1 = \ldots = k_m$ at first glance seems rather restrictive. However, if this condition is not assumed, then the following result indicates that the possibility of decoupling the composite system depends on the particular form of $B^*(t)$.

Suppose that $k^* = k_1 = \ldots = k_m$ does not hold and suppose, for simplicity, that $k^* = k_1 = \ldots = k_q$, $q < m$. Partition $\tilde{B}^*(t)$ as follows

$$\tilde{B}^*(t) = [\tilde{B}_1^*(t) \quad \tilde{B}_2^*(t)] \tag{20}$$

where $\tilde{B}_1^*(t)$ is m x q and $\tilde{B}_2^*(t)$ is m x (m-q).

Theorem 2 Suppose (as in Theorem 1) that the compensator (8) can be decoupled. If any q + 1 rows of $\tilde{B}_1^*(t)$ are nonzero for some $t\in[t_0,t_1]$, then the composite system can not be decoupled.

Since we are interested in compensators which can be used under general conditions, the compensator condition $k_1=\ldots=k_m$ is assumed throughout the remainder of the paper.

IV DECOUPLING WITH OUTPUT FEEDBACK

In this section the cascade compensator is used to allow the composite system to be decoupled using feedback of the compensator state z(t) and the plant output y(t). This approach is suggested by the results reported in [11] and [12]. In these references, a cascade compensator is used to allow pole placement and optimization by output feedback.

The results of this section are easily generalized to the case where there is a set of "available" state variables of the plant which is different than the set of outputs. Of course, the more state variables available for feedback, the lower the compensator order n_c which is required.

The following results hold for the composite system (9) under the assumption that $k_1 = k_2 = \dots = k_m = k^*$.

Lemma 5 For any nonnegative integer j,

$$\frac{d^j}{dt^j} y(t)=L_A^j(C)(t)x(t)+\sum_{q=0}^{j-1} L_D^q(L_A^{j-q-1}(C)(t)B(t)H(t))(t)z(t)$$
$$+\sum_{s=0}^{j-2}\frac{d^s}{dt^s}\Big[\sum_{q=0}^{j-2-s} L_D^q(L_A^{j-2-s-q}(C)(t)B(t)H(t))(t)E(t)r(t)\Big]. \qquad (21)$$

Theorem 3 Suppose that (A,C) is uniformly observable on $[t_0,t_1]$ with observability index ν_0 and suppose that both $B^*(t)$ in (4) and $E^*(t)$ in (17) are nonsingular at each $t\in[t_0,t_1]$. Then if the compensator is such that $k^* \geq \nu_0-d^*-2$, there exists a control law of the form

$$r(t) = \sum_{j=0}^{\nu_0-1} \hat{F}_{1j}(t)\frac{d^j}{dt^j} y(t)+\hat{F}_2(t)z(t)+\hat{G}(t)w(t), \qquad (22)$$

where $\hat{G}(t)$ is nonsingular at each $t\in[t_0,t_1]$, which decouples the system.

We have from Corollary 2 that, under the hypotheses of Theorem 3, $d^* \leq \nu_0-1$. Thus $\nu_0-d^*-2\in\{-1,0,1,2,\dots\}$.

Corollary 5 If for the given plant(1) the hypotheses of Theorem 3 hold and $\nu_0-d^*-2 = -1$, then no compensator is required for decoupling by feedback of the output and its time derivatives. If $\nu_0-d^*-2 = 0$, then a compensator of the form $\dot{z}(t)=r(t)$ can be used.

Under the requirements of Theorem 3 we have that (see Corollary 2)

$$\nu_0-d^*-2 \leq k^* \leq \frac{n_c-m}{m} \qquad (23)$$

Thus a lower bound on the order of the compensator in Theorem 3 is given by

$$n_c \geq m(\nu_0-d^*-1) \qquad (24)$$

In order to obtain a low order compensator, we must investigate relationships between k^* and n_c for different classes of systems. The particular plant must be considered since d^* enters into the requirements. However, there is one class of compensators which are easy to use for arbitrary plants. Let

$$D = \begin{bmatrix} 0 & I & 0 & \dots & 0 \\ 0 & 0 & I & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & I \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad E = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ I \end{bmatrix}, \quad H = [\, I \; 0 \; \dots \; 0 \,] \tag{25}$$

where I is the m x m identity matrix and 0 is the m x m zero matrix. If each row and column of D contains p submatrices, it is straightforward to calculate that $k_1 = k_2 = \dots = k_m = k^* = p-1$ and, of course, $n_c = mp$.

The form of compensator in (25) is not usually of minimal order but it has the advantage of applicability without regard to particular plant parameters. It also has the advantage that output derivatives are not required in the decoupling control law.

Theorem 4 Suppose (A,C) is uniformly observable on $[t_0 t_1]$ with observability index ν_0 and suppose $B^*(t)$ is nonsingular at each $t \in [t_0, t_1]$. Then if a compensator of the form (25) of order $n_c = m(\nu_0 - 1)$ is used, there exists a control law of the form

$$r(t) = F_1(t)y(t) + F_2(t)z(t) + \hat{G}(t)w(t) \tag{26}$$

which decouples the composite system.

It should be noted that these results can be generalized to compensators (25) which have, as the bottom row of D, nonzero m x m matrices.

V CONCLUSION

The problem of decoupling a time-variable plant using output feedback and a compensator has been considered. The solution of this problem is, in many ways, similar to the solution in the constant parameter case. However, the technical machinery needed to establish the results is much more complicated for the time-variable problem.

In this paper the problem formulation and the approaches to the solution have been chosen with considerable care in order to obtain rather general results without undue complexity. We shall cite two examples to illustrate this.

The assumption of constant d_i parameters can be relaxed with little mathematical difficulty. However, several practical problems arise which require careful consideration of the particular case at hand. For example, if $d_i = 1$ except for a small interval $[t_a, t_b)$ (or a single point t_c) where $d_i = 2$, the effect of coupling for such a length of time may be negligible. If so, this could save considerable complexity in the decoupling control law. Another difficulty that arises when d_i is nonconstant is that some gains (e.g. $G(t) = [B^*(t)]^{-1}$) grow without bound as the jump points are approached.

Secondly, compensators of lower order than those discussed here could be used in particular cases. However, we have opted for general applicability as illustrated by Theorems 1 and 2.

A major problem that remains is that of stabilizing the decoupled system in the time-variable case. It can be shown that under the restriction $m + \sum_{i=1}^{m} d_i = n$, the approach suggested by Wolovich can be used in this problem [13]. The results are that the decoupled system can be stabilized and, furthermore, this can be accomplished with plant output feedback. However, the restriction is quite severe and more general results are needed.

VI APPENDIX

Due to the lack of space, complete proofs cannot be given here but can be obtained from the authors. The following comments and proofs are meant to serve as an outline of the methods used. The argument (t) is dropped for convenience.

Lemmas 1,3,4, and 5 and the associated Corollaries are straightforward to prove by induction although the calculations are somewhat involved. Lemma 2 is proved by using Corollary 1 to show that the $m + \sum_{i=1}^{m} d_i$ row vectors

$$L_A^0(c_1),\ldots,L_A^{d_1}(c_1);\ldots;L_A^0(c_m),\ldots,L_A^{d_m}(c_m) \tag{27}$$

are linearly independent at each $t\in[t_o,t_1]$.

To prove Theorem 1, (12) and (14) are used to show that

$$\begin{aligned} &L_{\tilde{A}}^{j}(\tilde{c}_i)\tilde{B}=0, \quad j < d_i + k^* + 1 \\ &L_{\tilde{A}}^{d_i+k^*+1}(\tilde{c}_i)\tilde{B}= L_A^{d_i}(c_i)B\ \ L_D^{k^*}(H)E \\ &\qquad\qquad = L_A^{d_i}(c_i)BE^* \end{aligned} \tag{28}$$

The proof of Theorem 2 involves a somewhat detailed analysis of the particular form of (28).

Proof of Theorem 3 Using the definition of d* in (21) gives

$$\begin{aligned} \frac{d^j}{dt^j}\, y = L_A^j(C)x + \sum_{q=0}^{j-1} L_D^q\Big(L_A^{j-1-q}(C)BH\Big)z \\ + \sum_{s=0}^{j-2}\frac{d^s}{dt^s}\Big[\sum_{q=0}^{j-2-s-d^*} L_D^q\Big(L_A^{j-2-s-q}(C)BH\Big)Er\Big] \end{aligned} \tag{29}$$

By assumption $k^* \geq \nu_0 - d^* - 2$ and thus

$$L_D^q(H)E = 0, \qquad q = 0,1,\ldots,\nu_0 - d^* - 3 \tag{30}$$

Using this fact and (14) gives

$$L_D^q\left(L_A^{j-2-s-q}(C)BH\right)E = 0, \qquad \begin{array}{l} q = 0, 1, \ldots, \nu_0-d^*-3 \\ j = 0, 1, \ldots, \nu_0-1 \\ s = 0, 1, \ldots, \nu_0-3 \end{array} \tag{31}$$

Thus

$$\frac{d^j}{dt^j} y = L_A^j(C)x + \sum_{q=0}^{j-1} L_D^q\left(L_A^{j-1-q}(C)BH\right)z, \qquad j = 0, 1, \ldots, \nu_0-1$$

Substituting these expressions into (22), we must show that there exist matrices $\hat{F}_2$ and $\hat{F}_{1j}$, $j = 0, 1, \ldots, \nu_0-1$ such that the control law

$$r(t) = \sum_{j=0}^{\nu_0-1} \hat{F}_{1j}\left[L_A^j(C)x + \sum_{q=0}^{j-1} L_D^q\left(L_A^{j-1-q}(C)BH\right)z\right] + \hat{F}_2 z + \hat{G}w \tag{32}$$

decouples the composite system. However, by Theorem 1, there is a control law of the form (15) which decouples the system. Equating (15) and (32) gives

$$F_1 = \sum_{j=0}^{\nu_0-1} \hat{F}_{1j}\, L_A^j(C) \tag{33}$$

$$F_2 = \sum_{j=0}^{\nu_0-1} \hat{F}_{1j} \sum_{q=0}^{j-1} L_D^q\left(L_A^{j-1-q}(C)BH\right) + \hat{F}_2 \tag{34}$$

$$\tilde{G} = \hat{G}$$

Since (A,C) is uniformly observable with index ν_0, there is at least one set of $\hat{F}_{1j}$, $j=0,\ldots,\nu_0-1$ which satisfies (33). Then (34) gives $\hat{F}_2$. Thus we have a decoupling control law of the desired form.

To prove Theorem 4, let $u = u_1$ and set $\dot{u}_1 = u_2$, $\dot{u}_2 = u_3$, $\ldots$, $\dot{u}_{\nu_0-1} = r$. Then we can write

$$z = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{\nu_0-1} \end{bmatrix} \tag{35}$$

and show that $E^* = I$. Writing (32) for this case as

$$\dot{u}_{\nu_0-1} = \sum_{j=0}^{\nu_0-1} \hat{F}_{ij} \frac{d^j y}{dt^j} + \sum_{j=0}^{\nu_0-2} \hat{F}_{2j} \frac{d^j u_1}{dt^j} + \hat{G}w \tag{36}$$

and applying a transformation of the type discussed in [14,p357] gives a feedback control law with the desired properties.

VI REFERENCES

1. P.L. Falb and W.A. Wolovich, "Decoupling in the Design and Synthesis of Multivariable Control Systems", IEEE Trans. Automatic Control, AC-12 (1967), pp. 651-659.
2. E.G. Gilbert, "The Decoupling of Multivarialbe Systems by State Feedback", SIAM J. Control, 7 (1969), pp. 50-64.
3. B.S. Morgan, "The Synthesis of Linear Multivariable Systems by State Variable Feedback", Proc. 1964 JACC, Stanford, California, pp. 468-472.
4. Z.V. Rekasius, "Decoupling of Multivariable Systems by Means of State Variable Feedback", Proc. 3rd Ann. Allerton Conf. on Circuit and System Theory (Urbana, Ill, 1965), pp.439-447.
5. W.A. Porter, "Decoupling of and Inverses for Time-Varying Linear Systems", IEEE Trans. on Automatic Control, Vol.AC-14, No. 4, 1969, pp. 378-380.
6. W.J. Rugh, "On the Decoupling of Linear Time-Variable Systems", Proc. 6th Ann. Princeton Conf. on Information Processing and Systems, (Princeton, N.J.), 1971.
7. W.M. Wonham and A.S. Morse, "Decoupling and Pole Assignment in Linear Multivariable Systems: A Geometric Approach", SIAM J. Control, 8 (1970), pp. 1-18.
8. A.S. Morse and W.M. Wonham, "Decoupling and Pole Assignment by Dynamic Compensation", SIAM J. Control, 8 (1970), pp. 317-337.
9. J.W. Howze and J.B. Pearson, "Decoupling and Arbitrary Pole Placement in Linear Systems Using Output Feedback", IEEE Trans. Automatic Control! Vol. AC-15, No. 6, Dec. 1970, pp. 660-662.
10. L.M. Silverman and H.E. Meadows, "Controllability and Observability in Time-Variable Linear Systems", J. SIAM Control, Vol. 8, No. 1, 1967, pp. 64-73.
11. J.B. Pearson and C.Y. Ding, "Compensator Design for Multivariable Linear Systems", IEEE Trans. Automatic Control, Vol. AC-14, April 1969, pp. 130-134.
12. J.D. Ferguson and Z.V. Rekasius, "Optimal Linear Control Systems With Incomplete State Measurement", IEEE Trans. Automatic Control, Vol. AC-14, April 1969, pp. 135-140.
13. W.A. Wolovich, "On Stabilization of Controllable Systems", IEEE Trans. Automatic Control, Vol. AC-13, Oct. 1968, pp. 569-572.
14. L.A. Zadeh and C.A. Desoer, Linear System Theory, New York: McGraw-Hill, 1963,

POLE ASSIGNMENT BY MODE FEEDBACK*

H. HEMAMI
Department of Electrical Engineering
The Ohio State University
Columbus, Ohio 43210

ABSTRACT

The problem of pole assignment in linear systems by feedback is treated here by a method that utilizes the natural modes of the system. Explicit formulas are derived that describe the feedback gain for each mode as a function of present poles of the system and the desired poles. The approach is extended to systems with repeated eigenvalues, multiple input and multiple output as well as systems with output feedback.

The method requires knowledge of eigenvalues and eigenvectors of the system, and arose in study of biped and quadruped locomotion systems where linearized models of the dynamic system had to be thoroughly investigated; and, consequently, the eigenvalues and eigenvectors of the system were available.

This method may offer tutorial value where extensive knowledge of geometrical or algebraic structures is not available.

I. INTRODUCTION

Necessary and sufficient conditions for pole assignment for multi-input multioutput linear constant systems have been considered by Wonham[1], Davison[2], Wonham and Morse[3] and Wonham[4]. The problem of assigning all the poles of a completely controllable system has been considered in [1] and again [4] where triangularization and transformation to nth order scalar systems is used to prove that complete controllability is necessary and sufficient for pole assignment. An alternative proof is given in [2]. In [3], the pole assignment problem is treated from a geometrical point of view utilizing projection operators.

An algorithm,suitable for digital computers is developed by Bass and Gura[5] where explicit formulas for the feedback gains are derived in terms of powers of the system matrix and coefficients of the present and desired characteristic equations.

For systems with output feedback instead of state feedback two approaches are presented by Jameson[6], Davidson[7] and Nandi and Herzog[8].

In [6] and [8] it is proved that only poles corresponding to modes that are both controllable and observable can be assigned. In [6] an algorithm is described for assigning as many poles as there are output feedback, while in [7] an algorithm is described that, under certain conditions, assigns a number of poles very close to their desired values.

The present paper provides an alternative approach relying primarily on the availability or computation of eigenvalues and eigenvectors of the system. An objective comparison of this approach versus the ones described above from both analytical and computational point of view is beyond the scope of the present work.

II. STATE FEEDBACK

Consider the linear n-dimensional time-invariant system with p inputs, and q outputs

$$\begin{aligned} \overset{\circ}{X} &= FX + GU \\ Y &= HX \end{aligned} \qquad (1)$$

where F, G and H are compatible matrices. If eigenvalues, eigenvectors and

---

* This research was supported under NSF Grant number GK 25292.

dual basis vectors of F are available, the system can be diagonalized:

$$\begin{aligned} \overset{\circ}{Z} &= \Gamma Z + BU \\ Y &= CZ \end{aligned} \tag{2}$$

where vector Z describes the natural modes of the system. Several cases will be considered: (1) where Γ has distinct diagonal elements $\gamma_1, \gamma_2, \ldots,$ and γ_n; (2) where Γ is a block of Jordan form, namely, $\gamma_1 = \gamma_2 = \ldots = \gamma_n = \gamma$, and all elements immediately above the diagonal elements are 1's; and (3) where Γ consists of α blocks of Jordan forms of dimensions $\beta_1, \beta_2, \ldots \beta_\alpha$ $(\beta_1 + \beta_2 + \ldots + \beta_\alpha = n)$ respectively and eigenvalues $\gamma_1, \gamma_2, \ldots$ and γ_α respectively where $\gamma_1, \gamma_2, \ldots$ and γ_α are distinct.

Let p = q = 1, and let state feedback u = RZ be utilized. Let $\lambda_1, \lambda_2, \ldots$ and λ_n be desirable poles, which can be repeated or non-repeated. However, complex desirable poles should appear in conjugate pairs.

$$\det\,[\Gamma + BR - \lambda I] = \prod_{j=1}^{n} (\lambda_j - \lambda) \tag{3}$$

Factoring out $\Gamma - \lambda I$ and using results of [9] regarding a matrix identity one can show that Eq. (3) reduces to

$$[1 + R(\Gamma - \lambda I)^{-1}B] \prod_{j=1}^{n} (\gamma_j - \lambda) = \prod_{j=1}^{n} (\lambda_j - \lambda) \tag{4}$$

This in turn reduces to

$$r_i = \frac{1}{b_i} \, \frac{\displaystyle\prod_{j=1}^{n} (\lambda_j - \gamma_i)}{\displaystyle\prod_{\substack{j=1 \\ j\neq i}}^{n} (\gamma_j - \gamma_i)} \tag{5}$$

This equation provides the feedback gain r_i of the i^{th} mode in terms of present poles and the desired poles. Note that if $b_i=0$, the system is not controllable along the i^{th} mode, and hence γ_i cannot be changed. This means only poles corresponding to controllable modes can be assigned to desired locations. This equation complements Eq. (20) in reference [5].

This equation also renders solution to a dual problem: a linear system is completely observable through a scalar output; and it is desired to construct a control vector B such that all the poles are assigned. Alternately if a subset of modes are observable, only the corresponding poles can be assigned. This problem arises more naturally in control of locomotion systems where certain variables are easier to sense than others. These variables are outputs and would be used to control the system by design of vectors B.

Consider the second case where Γ is one block of Jordan form. Using the results in [9] again, Eq. (4) becomes

$$\left[1 + \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{(-1)^{j-i}}{(\gamma-\lambda)^{j-i+1}} \, r_i \, b_j\right] (\gamma-\lambda)^n = \prod_{j=1}^{n} (\lambda_j - \lambda) \tag{6}$$

If coefficients of powers of λ or $(\gamma-\lambda)$ are equated on both sides of Eq. (6) a system of linear equations in unknown vector R can be obtained which is lower triangular and consequently can be solved sequentially for r_1, r_2, ... and r_n:

$$r_1 = \frac{1}{b_n} (-1)^{n-1} \prod_{j=1}^{n} (\lambda_j-\gamma)$$

$$r_2 = \frac{1}{b_n} (-1)^{n-2} \left[\frac{b_{n-1}}{b_n} \prod_{j=1}^{n} (\lambda_j-\gamma) - \sum_{j=1}^{n} \prod_{\substack{i=1 \\ i\neq j}}^{n} (\lambda_i-\gamma) \right]$$

$$\cdots\cdots\cdots\cdots\cdots \qquad (7)$$

$$\cdots\cdots\cdots\cdots\cdots$$

Example 1: Let

$$\overset{\circ}{Z} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} Z + \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} u$$

Suppose the desired poles are -1, -2, -3 and -4. The feedback gain vector R from Eq. (7) is R = [-60, -47, 48, 46].

From Eq. (7) it is apparent that in order to assign all the poles the condition $b_n \neq 0$ must be satisfied. This means the system has to be completely controllable. If $b_n = 0$, one pole remains at γ. If $b_n = b_{n-1} = 0$ two poles remain at γ, etc.

Consider the third case where the system has α invariant subspaces of dimensions β_1, β_2, ... and β_α. Let the components of vector B be relabeled as follows: b_1^1, b_2^1, ... $b_{\beta_1}^1$ are the first β_1 elements of B followed by b_1^2, b_2^2, ... $b_{\beta_2}^2$, ..., and finally b_1^α, b_2^α, ..., $b_{\beta_\alpha}^\alpha$. Similarly let $R = [R^1, R^2, \ldots, R^\alpha]$. Eq. (4) still holds. With the above relabeling of B and R, it reduces to

$$\left[1 + \sum_{k=1}^{\alpha} \sum_{i=1}^{\beta_\alpha} \sum_{j=i}^{\beta_\alpha} \frac{(-1)^{j-i}}{(\gamma_k-\lambda)^{j-i+1}} r_i^k b_j^k\right] \prod_{k=1}^{\alpha} (\gamma_k-\lambda)^{\beta_k} = \prod_{j=1}^{n} (\lambda_j-\lambda) \qquad (8)$$

This equation is linear in $R^k (k = 1, 2, \ldots, \alpha)$ and for given Λ and B can be solved for R. The approach is similar to the case with one Jordan block case. Again it can be shown that $b_{\beta_k}^k$ $(k = 1, 2, \ldots, \alpha)$ has to be nonzero so that all the poles may be assigned.

This approach can be similarly applied to systems where Γ has other Jordan forms. For such cases Eq. (4) still holds.

Example 2: Consider the following system which is not completely controllable with one input

$$\overset{\circ}{Z} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} Z + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u$$

Suppose the desired poles are -1, -1 and -1. For this system Eq. (4) becomes

$$(1-\lambda)\left[(1-\lambda)^2 - r_1 + (r_2+r_3)(1-\lambda)\right] = (-1-\lambda)^3$$

This equation cannot be satisfied because the left side contains $(1-\lambda)$ as a factor and consequently one pole cannot be assigned. The other two poles can be assigned to -1 with $r_1 = -4$ and $r_2 + r_3 = -4$.

Now consider state feedback for a multi-input system. The control vector is $U = RZ$ where row R is $p \times n$. Consider the distinct eigenvalue case first. If the i^{th} row of B consists of all zeroes, the i^{th} mode is not controllable and obviously γ_i cannot be changed. If some elements of the i^{th} row are nonzero the possibility exists of assigning γ_i by any one of the inputs that control that mode or by several inputs: sequentially by assigning γ_i to several intermediate positions before it finally is assigned to λ_i.

Example 3: Let the desired poles be (-1, -2, -3, -4) for

$$\overset{\circ}{Z} = \left|\begin{matrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & +2 \end{matrix}\right| Z + \left|\begin{matrix} 1 & 2 \\ 2 & -1 \\ 1 & -1 \\ 2 & 1 \end{matrix}\right| \begin{matrix} u_1 \\ u_2 \end{matrix}$$

This system is completely controllable via input u_1 and one can just use one input to assign all the poles:

$$R_1 = \left|\begin{matrix} 0 & 0 & 20 & -15 \\ 0 & 0 & 0 & 0 \end{matrix}\right|$$

Alternately one can use the second input for the same purpose:

$$R_2 = \left|\begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -20 & -30 \end{matrix}\right|$$

Now suppose the first input is used

$$u_1 = \left[190/11 \quad -193/40 \quad 190/11 \quad -190/11\right] Z$$

The poles of the system are transferred to intermediate positions from which they are transferred to the desired positions by

$$u_2 = \left[-20/11 \quad 7/20 \quad -20/11 \quad 20/11\right] Z$$

Consequently

$$R_3 = \left[\begin{matrix} 190/11 & -193/40 & 190/11 & -190/11 \\ -20/11 & 7/20 & -20/11 & 20/11 \end{matrix}\right]$$

is a third solution. There are many other solutions. To carry the pole assignment for the above either of two approaches can be followed: (1) use the inputs sequentially where for every scalar input state feedback is applied and a set of n intermediate poles are assigned, or (2) parameterize the controls in one scalar control: $U = Su$ where vector S is $p \times 1$ and arbitrary, but should not be orthogonal to any row vectors of B. For vector u one derives feedback gains according to Eq. (5). Let this feedback vector be R_u, then $R = SR_u$.

For systems with repeated eigenvalues either the sequential approach or the parameterization approach is applicable (with proper modification of the equations), but parameterization may destroy controllability and consequently should not be used.

Example 4: For the following system parameterization destroys controllability

$$\overset{\circ}{Z} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} Z + \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{matrix} u_1 \\ u_2 \end{matrix}$$

In the above discussion the input vector U was parameterized to one scalar input. The possibility of reducing the size of U to any size between 1 and p by parameterization always exists.

Eqs. (5) and (7) describe the feedback gains in the canonical coordinate system. These equations can also be translated in the original coordinate system so that computation of eigenvectors are not necessary. This is demonstrated below for a system with distinct eigenvalues. Let p = q = 1. The system equations are

$$\begin{aligned} \overset{\circ}{X} &= FX + Gu \\ u &= OX \end{aligned} \qquad (9)$$

where vector O is 1 x n. Let T be a vector with components

$$t_i = \frac{\prod_{j=1}^{n} (\lambda_j - \gamma_i)}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (\lambda_j - \gamma_i)} \qquad (10)$$

Let P_1, P_2, ... and P_n be one-dimensional projection operators associated with F:

$$F = \sum_{i=1}^{n} \gamma_i P_i \qquad (11)$$

It can be shown that

$$OP_iG = t_i \qquad i = 1, \ldots, n \qquad (12)$$

or finally

$$\begin{aligned} OG &= \sum_{i=1}^{n} \gamma_i t_i \\ OFG &= \sum_{i=1}^{n} \gamma_i^2 t_i \\ &\cdots\cdots\cdots \\ OF^{n-1}G &= \sum_{i=1}^{n} \gamma_i^n t_i \end{aligned} \qquad (13)$$

Eq. (13) is a set of linear equations which can be solved for O when G, γ_i,

λ_i are given. Alternately if O is given. one can solve for G.

III. OUTPUT FEEDBACK

Suppose the original system has distinct eigenvalues. Consider first the inverse problem of pole assignment.

Suppose vectors R and B in Eq. (5) are given and the objective is to find the new poles Λ of the system. Solution of the system of nonlinear equations

$$r_i b_i = t_i(\Lambda, \Gamma) \tag{14}$$

for Λ is not easy and one may have to resort to numerical solutions. One interesting result that can be exploited, in this connection, is that if Newton-Raphson method is used, the inverse of the Jacobian matrix $\frac{\partial T}{\partial \Lambda}$ can be analytically derived

$$\left[\frac{\partial T}{\partial \Lambda}(\Lambda, \Gamma)\right]^{-1} = \frac{\partial T}{\partial \Lambda}(\Gamma, \Lambda) \tag{15}$$

In other words interchange of γ_i for λ_i and λ_i for γ_i $(i = 1, 2, \ldots, n)$ in $\frac{\partial T}{\partial \Lambda}$ renders its inverse.

Now consider the system with p inputs and q outputs with distinct eigenvalues. Assume in each row of B at least one number is nonzero so that the system is completely controllable. By parameterization reduce the system to one input u(t) where p x 1 vector S is arbitrary. For this system use Eq. (5) to derive a n x 1 vector R. The question is whether R can be written as a linear sum of rows of C: R = LC where L is 1 x q and C is q x n. It is apparent that if any column j of C is all zeroes, for any L, r_j is zero and consequently γ_j cannot be assigned. On the other hand, column j of C being all zeroes means that mode j of the system is not observable. Thus if a mode is not observable, the corresponding pole cannot be assigned. This proof can easily be extended to input and output parameterizations of higher order: Let K be p x q and let $\Gamma + BKC$ be the matrix of the system with output feedback. If row i of B or column j of C are zero γ_i or γ_j is, respectively, a permanent pole of the system. Now suppose the system is completely controllable and completely observable. Also assume the $q(q < n)$ rows of C are independent. One method of output feedback design is to parameterize both the input and the output: U = Su and y = LY. Let BS be a column vector W and LC be a row vector V where S and L are unknown p x 1 and 1 x q vectors respectively. The pole assignment problem is to derive S and L such that

$$\begin{aligned} BS &= W \\ LC &= V \\ w_1 v_1 &= t_1(\Lambda, \Gamma) \\ w_2 v_2 &= t_2(\Lambda, \Gamma) \\ &\cdots\cdots \\ w_n v_n &= t_n(\Lambda, \Gamma) \end{aligned} \tag{16}$$

This system of equations is quadratic in S and L and may or may not have a solution. The latter n equations involve p + q variables s_i and ℓ_i. If $p + q \geq n$ in general a solution may exist because there are more variables than equations.

In solution of these nonlinear equations by numerical methods Eq. (15) can be exploited for the needed inverse.

If S is selected arbitrarily, these equations are linear in L and q poles could be assigned. If L is selected arbitrarily and rows of B are independent P poles could be assigned. For other approaches to this problem see [6] and [7].

Example 5: Consider the example in [7]

$$\overset{\circ}{Z} = \begin{vmatrix} -1 & 0 \\ 0 & -4 \end{vmatrix} Z + \begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix} U$$

$$Y = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} X$$

with the above approach one obtains

$$\ell_1 s_2 = -27$$

$$\ell_2 s_1 = 12$$

where two of the variables are arbitrary, and hence many solutions exist. For this particular case one can actually assign additional constraints on L and S such that some norm of the gain SL is minimized. In this example $p = q = n$ and columns of B and rows of C are independent. One can diagonalize and assign all the poles by

$$U = KY$$

$$K = \begin{bmatrix} 0 & -6 \\ -9 & 0 \end{bmatrix}$$

These gains are lower than the given values in [7] by a factor of 5.

Higher order parameterizations can also by considered.

SUMMARY

The problem of pole assignment by feedback has been considered within an elementary framework. The natural modes of the system are used in deriving state feedback gains for systems with repeated or nonrepeated eigenvalues. The results can also be expressed in the original coordinate system. Parameterization is used to consider multi-input - multioutput systems as well as output feedback.

REFERENCES

1. W. M. Wonham, "On pole assignment in multi-input, controllable linear systems," IEEE Trans. Automatic Control, Vol. AC-12, pp. 660-665, December 1967.

2. E. J. Davison and W. M. Wonham, "On pole assignment in multi-variable linear systems," IEEE Trans. Automatic Control (Correspondence), Vol. AC-13, pp. 747-749, December 1968.

3. A. S. Morse and W. M. Wonham, "Decoupling and pole assignment by dynamic compensation," NASA Tech. Rept. ERC PM-86, July 1969; also SIAM J. Control, Vol. 8, August 1970.

4. A. S. Morse and W. M. Wonham, "Triangular decoupling of linear multivariable systems," IEEE Trans. Automatic Control, Vol. AC-15, pp. 447-449, August 1970.

5. R. W. Bass and I, Gura, "High order system design via state space considerations," Reprints, Joint Automatic Control Conference, Troy, New York, pp. 311-318, June 1965.

6. A. Jameson, "Design of a single input system for specified roots using output feedback," IEEE Trans. Automatic Control, Vol. AC-15, pp. 345-348, June 1970.

7. E. J. Davison, "On pole assignment in linear systems with incomplete state feedback," IEEE Trans. Automatic Control, Vol. AC-15, pp. 348-351, June 1970.

8. A. K. Nandi and J. H. Herzog, "Comments on design of a single input system for specified roots using output feedback," IEEE Trans. Automatic Control, Vol. AC-16, pp. 384-385, August 1971.

9. H. Hemami, "Derivation of a matrix identity," IEEE Trans. Automatic Control, Vol. AC-16, pp. 303-304, June 1969.

A NEW ALGORITHM FOR THE IRREDUCIBLE REALIZATION OF A RATIONAL MATRIX

CLAUDIO GORI GIORGI and ALBERTO ISIDORI
Istituto di Automatica, Università di Roma, Rome, Italy

ABSTRACT

This paper deals with the problem of constructing a minimal realization of a rational matrix, by reducing a non-minimal one. The proposed reduction procedure is a new one, as far as the authors know, and it has the advantage of requiring only a limited number of operations on the matrices characterizing the non-minimal realization. Therefore both the dimension of the minimal realization and the minimal realization itself can be computed with a high degree of precision.

1. INTRODUCTION

The problem of computing minimal realizations of a rational matrix has been widely considered by several authors. The existing procedures can be classified in different ways, according to the mathematical techniques and to the particular algorithms involved. One can distinguish between procedures for the direct computation of minimal realizations and procedures based upon the reduction of non-minimal ones; from another point of view it is possible to classify the methods on the basis of the nature of the initial data (matrix of rational proper functions, partial fraction expansion, etc.) as well as of that of the terminal data (Jordan form, block-companion form, etc.).

With reference mainly to the algorithms involved, there are some methods based on the manipulation of polynomial matrices [5,8,9,12] (most of them require the construction of McMillan-Smith forms); others give the realization in Jordan canonical form [3,4,10,11]; others require operations that in any case are equivalent to the evaluation of the transition matrix or, what is the same, of a suitable number of powers of the dynamic matrix [1,2,6,7,14]; lastly, in [13], an essentially different procedure has been proposed, which is based on performing elementary row and column operations directly on the matrices characterizing a non-minimal realization.

Some of these methods require first the computation of a non-minimal realization; this can be easily performed following the indications given in [7,page 181, Method B]. In this way one obtains a realization which, according to the particular case, is certainly controllable or observable and this obviously simplifies the subsequent reduction operations.

In this paper we propose a new method for computing minimal realizations, which is substantially different from all the methods listed above.

Let us consider a qxp proper rational matrix

$$W(s) = \sum_{i=1}^{l} \frac{R_i(s)}{(s-\lambda_i)^{\mu_i}} \tag{1}$$

where $R_i(s)$, $i=1,\dots,l$, are polynomial matrices. As is known[8], the direct sum of the realizations (A_i, B_i, C_i), $i=1,\dots,l$, of each term $\frac{R_i(s)}{(s-\lambda_i)^{\mu_i}}$ of (1) is a realization of W(s). Moreover, if each realization (A_i, B_i, C_i) is minimal (i.e. controllable and observable), the realization of W(s) obtained by direct sum is also minimal. In what follows we will consider, therefore, the problem of constructing a minimal realization of a rational matrix of the type $\frac{R(s)}{(s-\lambda)^{\mu}}$. The minimal realization will be obtained through the reduction of a given non-minimal one which, as we have already observed, can be constructed very easily; let the latter be characterized by the triplet of matrices (A, B,C), respectively nxn ($n=\min[\mu p, \mu q]$), nxp, qxn.

In what follows we will be concerned with the procedure for the reduction of the triplet (A,B,C) to a controllable triplet, since the procedure based on the reduction to an observable one (which can be used alternatively or successively) is obviously dual.

2. PRINCIPLE OF THE PRESENT METHOD

It is well known [7] that a triplet of matrices (A,B,C) is non-controllable if and only if there exists a nonsingular nxn matrix T such that

$$\tilde{A} \triangleq TAT^{-1} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \qquad \tilde{B} \triangleq TB = \begin{bmatrix} 0 \\ \tilde{B}_2 \end{bmatrix} \tag{2}$$

where $\tilde{A}_{22}$ and $\tilde{B}_2$ are respectively (n-r)x(n-r) and (n-r)xp. If T is partitioned in the form

$$T = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \tag{3}$$

where T_1 is rxn, eqs.(2) are equivalent to

$$T_1 A = \tilde{A}_{11} T_1 \tag{4a}$$

$$T_1 B = 0$$

and

$$\begin{aligned} T_2A &= \tilde{A}_{21}T_1 + \tilde{A}_{22}T_2 \\ T_2B &= \tilde{B}_2 \end{aligned} \tag{4b}$$

In order to reduce (4a) to a form which is more suitable for our subsequent developements, we observe that, since $\tilde{A}$ is similar to A and - by hypothesis - $\det(sI-A) = (s-\lambda)^n$, one can assume

$$\tilde{A}_{11} = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ \tilde{a}_{21} & \lambda & \dots & 0 \\ \cdot & \cdot & \dots & \cdot \\ \tilde{a}_{r1} & \tilde{a}_{r2} & \dots & \lambda \end{bmatrix} \tag{5}$$

As a consequence of (5), eqs. (4a) can be satisfied if and only if the system

$$\begin{aligned} &X(A-\lambda I \quad B) = (YX \quad 0) \\ &y_{ij} = 0 \qquad (i \le j) \end{aligned} \tag{6}$$

in the unknown matrices X and Y, respectively rxn and rxr, has at least one solution $(\overline{X},\overline{Y})$ with $\overline{X}$ of full rank. Note that system (6), as a consequence of the assumption on $\tilde{A}_{11}$, is triangular (i.e., the i-th rows of X and Y can be calculated from the values of the preceding i-1 rows). To take into consideration the dimensions of the unknown matrices, system (6) will henceforth be denoted as system $\Sigma(r)$.

One can easily verify that, once such a solution of system (6) has been calculated, the reduction of the triplet (A,B,C) is quite strightforward. In fact, if one assumes $T_1=\overline{X}$ (and also $\tilde{A}_{11}=\overline{Y} + \lambda I$, but the latter is of no interest in the reduction) in eqs. (4a), there remain to be solved only eqs. (4b). Their excess of unknowns is (n-r)xn; consequently T_2 can be arbitrarily chosen (see section 3) provided that $\det(T) \neq 0$. Once a matrix T satisfying (3) has been completly determined, the triplet (A, B,C) can be reduced to the triplet $(\tilde{A}_{22},\tilde{B}_2,\tilde{C}_2)$ by the relations

$$\begin{aligned} \tilde{A}_{22} &= \text{last } n-r \text{ columns of } T_2AT^{-1} \\ \tilde{B}_2 &= T_2B \\ \tilde{C}_2 &= \text{last } n-r \text{ columns of } CT^{-1} \end{aligned} \tag{7}$$

Let r_o be the largest value of r such that $\Sigma(r)$ has at least one solution $(\overline{X},\overline{Y})$ with $\overline{X}$ of full rank; one can immediately verify that the above procedure, if applied with $r=r_o$, must supply a controllable triplet of matrices $(\tilde{A}_{22},\tilde{B}_2,\tilde{C}_2)$.

From all the above considerations, we deduce that a possible way to reduce the triplet (A,B,C) to a controllable triplet is the following:

(i) find the value of $r_\circ$,
(ii) find a solution (X,Y) of $\Sigma(r_\circ)$ such that X has full rank,
(iii) apply relations (7) (provided T_2 has been properly chosen)

In the following section we will describe an algorithm which allows a simultaneous solution of the problems stated in (i) and (ii) and we will give indications on how to perform the operations implied by (iii).

3. REDUCTION ALGORITHM

Since system $\Sigma(r)$ is triangular, it seems natural to find the solution recurrently; this corresponds to solving successively systems $\Sigma(1), \Sigma(2), \dots$. The algorithm that we will now describe is one of the possible procedures that can be followed for the actual evaluation of these solutions; it also makes it possible, as we will show later, to find the unknown $r_\circ$ and, therefore, the solutions of the problems stated in (i) and (ii).

ALGORITHM - *Consider the* nx(n+p+n) *matrix*

$$K_1 = (H_1 \quad I) \tag{8}$$

where $H_1 = (A-\lambda I \quad B)$ *and* I *is an* nxn *identity matrix. Perform row operations to triangularize* H_1; *if* L_1 *is the matrix which performs these operations and if* $H_2=L_1H_1$, *one has*

$$L_1K_1 = (H_2 \quad L_1) \triangleq K_2 \tag{9}$$

If $rank(H_1)=n-m_1$ $(m_1 \geq 0)$, *the last* m_1 *rows of* H_2 *are zero. Add to* K_2 *the following matrix*

$$\begin{pmatrix} 0 & 0 & 0 \\ L_1^{(m_1)} & 0 & -L_1^{(m_1)} \end{pmatrix} \tag{10}$$

where $L_1^{(m_1)}$ *is the submatrix of* L_1 *formed from the last* m_1 *rows. Let* $\overline{K}_2 = (\overline{H}_2 \quad \overline{L}_1)$ *be the resulting matrix* ($\overline{H}_2, \overline{L}_1$ *have, respectively, the dimensions of* H_2, L_1). *Repeat on* $\overline{K}_2$ *the same procedure as the one performed on* K_1 *and so on. At the step* r *consider the matrix* $\overline{K}_r = (\overline{H}_r \quad \overline{L}_{r-1})$ *and triangularize* $\overline{H}_r$ *with row operations performed by the matrix* L_r, *obtaining the matrix*

$$L_r\overline{K}_r = (H_{r+1} \quad L_r\overline{L}_{r-1}) \triangleq K_{r+1} \tag{11}$$

where the last m_r *rows of* $H_{r+1} = L_r\overline{H}_r$ *are zero; add to* K_{r+1}

the matrix

$$\begin{pmatrix} 0 & 0 & 0 \\ L_r^{(m_r)} & 0 & -L_r^{(m_r)} \end{pmatrix} \tag{12}$$

where $L_r^{(m_r)}$ *is formed from the last* m_r *rows of* $L_r\bar{L}_{r-1}$ *, thus obtaining the matrix* $\bar{K}_{r+1}$.

In the next section it will be proved that:

(a) There exists an integer $s\leq n$ such that $\operatorname{rank}(\bar{H}_s)=n$. Therefore the procedure terminates in at most n steps.

(b) The quantity $m=\sum_{i=1}^{s-1} m_i$ (which is $<n$) is equal to the above-defined quantity $r_\circ$.

(c) The mxn matrix

$$L = \begin{pmatrix} L_1^{(m_1)} \\ L_2^{(m_2)} \\ \vdots \\ L_{s-1}^{(m_{s-1})} \end{pmatrix} \tag{13}$$

has full rank and there exists a matrix $\bar{Y}$ such that $(L,\bar{Y})$ is a solution of $\Sigma(m)$.

Following the principle described in section 2, in order to reduce the triplet (A,B,C) we can therefore assume $T_1=L$ and then proceed by applying (7), with $r=r_\circ$.

The matrix T_2, which is arbitrary, can be conveniently chosen in order to reduce computations. In particular, if P is an nxn permutation matrix such that the first $r_\circ$ columns of LP give a nonzero minor, one can choose T_2 so that

$$\begin{pmatrix} L \\ T_2 \end{pmatrix} P = \begin{pmatrix} \hat{T}_{11} & \hat{T}_{12} \\ 0 & I \end{pmatrix} \tag{14}$$

As a consequence, by putting

$$P^{-1}AP = \begin{pmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{pmatrix}, \quad P^{-1}B = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix}, \quad CP = (\hat{C}_1 \;\; \hat{C}_2) \tag{15}$$

one has

$$\begin{aligned} \tilde{A}_{22} &= \hat{A}_{22} - \hat{A}_{21}\hat{T}_{11}^{-1}\hat{T}_{12} \\ \tilde{B}_2 &= \hat{B}_2 \\ \tilde{C}_2 &= \hat{C}_2 - \hat{C}_1\hat{T}_{11}^{-1}\hat{T}_{12} \end{aligned} \tag{16}$$

It is useful to stress that the proposed algorithm for reducing the realization (A,B,C) to the controllable one $(\tilde{A}_{22},\tilde{B}_2,\tilde{C}_2)$ involves a fairly limited number of operations. In fact, when the matrix L is determined, after the initial triangularization of H_1, in the remaining triangularization of matrices $\bar{H}_r$ only the last m_r rows are involved. This is equal, in any case, to the triangularizations, with row operations, of a single $(n+r_\circ)x(2n+p)$ matrix, stopping after n steps. Then is necessary to find and invert the matrix $\hat{T}_{11}$, which is $r_\circ x r_\circ$ and, lastly, the products and the differences indicated in (16) must be performed.

4. PROOF OF THE REDUCTION ALGORITHM

We will first of all verify that the aforementioned algorithm corresponds to finding a particular solution of systems $\Sigma(m_1)$, $\Sigma(m_1+m_2)$,... . For this purpose we consider the system $\Sigma(1)$

$$x(A-\lambda I \quad B) = 0 \tag{17}$$

where x is a 1xn row-vector. If $(A-\lambda I \quad B)$ has rank n the algorithm terminates at the first step, with $m_1=0$; then, too, (17) admits only the solution x=0 and, therefore, $r_\circ=0$ by definition. If $(A-\lambda I \quad B)$ has rank $n-m_1$,(17) admits m_1 linearly independent solutions and it is easy to verify that the m_1 rows of $L_1^{(m_1)}$ can be taken as such; by hypothesis the last m_1 rows of $L_1(A-\lambda I \quad B)$ are zero and, therefore, (17) is satisfied by m_1 linearly independent row-vectors of the type

$$e_{n-m_1-i}L_1 = x_i \qquad (i=1,...,m_1) \tag{18}$$

where e_j is a 1xn row-vector with the element (1,j) equal to 1 and the remaining equal to 0. As a consequence of this we can conclude that system $\Sigma(m_1)$ admits the solution $X=L_1^{(m_1)}$, Y=0; moreover the solution for X has full rank.

Consider now the system $\Sigma(m_1+1)$; the latter, thanks to the property of triangularity, is equivalent to system $\Sigma(m_1)$, solved as indicated above, and to the further equation

$$(x_{m_1+1} \quad y_{m_1+1}) \cdot \begin{pmatrix} A-\lambda I & B \\ L_1^{(m_1)} & 0 \end{pmatrix} = 0 \tag{19}$$

where x_{m_1+1} is a 1xn row-vector, the last row of the unknown X, and $y_{m_1+1} \triangleq (-y_{m_1+1,1} \cdots -y_{m_1+1,m_1})$. It is easy to verify that the coefficient matrix of (19), which is $(n+m\ )x(n+p)$, has rank $n-m_2$ $(m_2 \leq m_1)$; as a consequence, (19) admits m_1+m_2 linearly independent solutions. m_1 of the latter can be of the type $(x_i \quad 0)$, $i=1,...,m$, where x_i denotes the i-th row of $L_1^{(m_1)}$, a solution of (17). The remaining m_2 can be calculated by means of a further triangularization. This operation can be performed by premultiplying the coefficient matrix of (19) by the matrix

$$\begin{pmatrix} L_2 & 0 \\ 0 & I \end{pmatrix} P \begin{pmatrix} L_1 & 0 \\ 0 & I \end{pmatrix} \tag{20}$$

where the identity matrices are $m_1 x m_1$ and P is a matrix that permutes the block of the last m_1 rows with the block immediately above (the first n rows of the matrix thus triangularized form a matrix equal to the above-defined matrix H_3 and the submatrix formed from the first n rows and n columns of (20) is equal to the matrix L_2L_1). The last m_1+m_2 rows of (20) are linearly independent solutions on (19) and they have the form

$$\begin{pmatrix} L_2^{(m_2)} & \bar{Y} \\ L_1^{(m_1)} & 0 \end{pmatrix} \tag{21}$$

where $\bar{Y}$ is the submatrix formed from the last m_2 rows and m_1 columns of L_2. Consequently we can conclude that the system $\Sigma(m_1+m_2)$ admits the solution

$$X = \begin{pmatrix} L_1^{(m_1)} \\ L_2^{(m_2)} \end{pmatrix} \qquad y_{ij} = \begin{cases} \bar{y}_{ij} & \begin{cases} i=m_1+1,\dots,m_1+m_2 \\ j=1,\dots,m_1 \end{cases} \\ 0 & \text{everywhere else} \end{cases} \tag{22}$$

By proceeding in this way we can identify the r-th step of the algorithm with the search for new solutions (m_r in number) for the system $\Sigma(m_1+m_2+\dots+m_r)$, in addition to the $m_1+m_2+\dots+m_{r-1}$ solutions calculated at the previous steps. In fact we determine only the solution corresponding to the unknown X, which has the form

$$\begin{pmatrix} L_1^{(m_1)} \\ L_2^{(m_2)} \\ \vdots \\ L_r^{(m_r)} \end{pmatrix} \tag{23}$$

and we disregard Y, which is of no interest in our reduction procedure.

If at some step s we have $m_s=0$, it is not possible to find further solutions of $\Sigma(m_1+m_2+\dots+m_{s-1}+1)$ linearly independent from the ones calculated for $\Sigma(m_1+m_2+\dots+m_{s-1})$: this corresponds to the end of the algorithm owing to the condition $\text{rank}(\bar{H}_s)=n$.

We have now to prove the assertions (a),(b) and (c).

As regards (c), it is easy to verify that L, having the form (23) is a solution for X of $\Sigma(m)$, corresponding to a suitable value $\tilde{Y}$ of Y; moreover L has rank m thanks to a result reported in the Appendix, since $(L\ \tilde{Y})$ has rank m by construction.

As regards (b), we have only to prove that the end of the procedure due to the impossibility of finding a solution (X,Y) of $\Sigma(m+1)$ with X of full rank does not depend on the particular solution previously found for $\Sigma(m)$. For this purpose we observe that, if T_2 is any (n-m)xn matrix such that $T = (L' \; T_2')'$ is nonsingular (' denotes transposition), we have by construction

$$TAT^{-1} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \qquad TB = \begin{bmatrix} 0 \\ \tilde{B}_2 \end{bmatrix} \tag{24}$$

where $\tilde{A}_{22}$ is (n-m)x(n-m) and $\tilde{B}_2$ is (n-m)xp. If the pair $(\tilde{A}_{22},\tilde{B}_2)$ were not controllable (that is if $m<r_\circ$), there should exists a nonsingular (n-m)x(n-m) matrix $\tilde{T}$ such that

$$\tilde{T}\tilde{A}_{22}\tilde{T}^{-1} = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ x & x & \dots & x \\ . & . & \dots & . \\ x & x & \dots & x \end{bmatrix} \qquad \tilde{T}\tilde{B} = \begin{bmatrix} 0 & \dots & 0 \\ x & \dots & x \\ . & \dots & . \\ x & \dots & x \end{bmatrix} \tag{25}$$

where x denotes possibly nonzero elements. Consider now the matrix

$$F = \begin{bmatrix} I & 0 \\ 0 & \tilde{T} \end{bmatrix} T \tag{26}$$

where I is mxm. Clearly the first m rows of F are equal to L; moreover, thanks to (24) and (25), the first m+1 rows of F are a solution for X of $\Sigma(m+1)$ (corresponding to a suitable solution for Y). The latter fact would imply that $\Sigma(m+1)$ should admit at least one solution (X,Y) with X of full rank and whose first m rows form exactly the particular solution previously found for $\Sigma(m)$, but this contradicts the hypothesis. Consequently $(\tilde{A}_{22},\tilde{B}_2)$ is controllable and, therefore, $m=r_\circ$.

The proof of (a), lastly, is trivial. Since the stop condition ($\text{rank}(\bar{H}_s)=n$) is equivalent to $m_s=0$, as we have already observed and since n is finite, the procedure terminates in a finite number of steps.

5. CONCLUSIONS

A new algorithm for computing minimal realizations of a rational matrix has been proposed; this algorithm operates by reducing a suitable non-minimal realization of the given rational matrix. The proposed procedure requires a fairly limited number of operations on the initial data, above all if the amount of the reduction is small compared with the dimension of the non-reduced realization. This obviously implies a reduction of the computing time as well as of the numerical errors made in the course of the application of the algorithm.

6. ACKNOWLEDGMENTS

The authors wish to thank Professor A.Ruberti who suggested the topic of this research.

APPENDIX

LEMMA - *Consider the system*

$$XH = (YX \quad 0)$$
$$y_{ij} = 0 \qquad i \leq j \tag{A.1}$$

where the coefficient matrix H *is* nxs $(s > n)$ *, the unknown matrices* X *and* Y *are respectively* mxn *and* mxm $(m \leq n)$ *; if* X,Y *are a solution of* (A.1) *, then*

$$\text{rank}(X \quad Y) = m \quad <=> \quad \text{rank}(X) = m \tag{A.2}$$

Proof. The implication <= is trivial. As regards the implication =>, the proof will be carried out by induction. Let $\Sigma(m)$ denote the system (A.1), let X(m) and Y(m) denote any solution of $\Sigma(m)$ and suppose that (A.2) is true (for the actual value of m). Further, let $\bar{X}(m+1)$,$\bar{Y}(m+1)$ be a solution of $\Sigma(m+1)$; by construction the first m rows of these matrices give a particular solution of system $\Sigma(m)$. This solution will be denoted by $\bar{X}(m)$, $\bar{Y}(m)$. If

$$\text{rank}(\bar{X}(m+1) \quad \bar{Y}(m+1)) = m+1 \tag{A.3}$$

it follows that $\text{rank}(\bar{X}(m) \ \bar{Y}(m)) = m$ and, therefore,

$$\text{rank}(\bar{X}(m)) = m \tag{A.4}$$

because (A.2) is supposed to hold, by hypothesis, for $\Sigma(m)$. Assume now that $\text{rank}(\bar{X}(m+1)) \neq m+1$; as a consequence, there should exist a 1x(m+1) nonzero row-vector γ, such that $\gamma\bar{X}(m+1) = 0$. Postmultiplying the latter by H, one obtains

$$0 = \gamma\bar{X}(m+1)H = \gamma(\bar{Y}(m+1)\bar{X}(m+1) \quad 0) = \gamma\bar{Y}(m+1)(\bar{X}(m+1) \quad 0) \tag{A.5}$$

Owing to the form of $\bar{Y}(m+1)$, the element in the (1,m+1) position of $\gamma\bar{Y}(m+1)$ is zero; therefore, (A.4) implies

$$\gamma\bar{Y}(m+1) = 0 \tag{A.6}$$

From the latter and from the definition of γ it follows that

$$\gamma(\bar{X}(m+1) \quad \bar{Y}(m+1)) = 0 \tag{A.7}$$

which contradicts the hypothesis (A.3). The proof by induction is therefore completed, since (A.2) is trivial in the case m=1.

REFERENCES

1. N.D.ALBERTSON and B.F.WOMACK, An algorithm for constructing irreducible realizations of time-invariant systems, J. of the Franklin Inst., 285(1968), pp. 110-124.

2. C.BRUNI, A.ISIDORI and A.RUBERTI, A method of realization based on the moments of the impulse-response matrix, IEEE

Trans. on Aut. Control., 14 (1969), pp.203-4.

3. E.G.GILBERT, Controllability and observability in multivariable control systems, J.SIAM Control, 1 (1963), pp.128-151.
4. C.GUEGUEN and E.TOUMIRE, Mise sous forme de variable d'état par la méthode matricielle, Automatisme, 13(1968), pp.199-203
5. M.HEYMANN and J.A.THORPE, Transfer equivalence of linear dynamical systems, J.SIAM Control, 8 (1970), pp. 19-40.
6. B.L.HO and R.E.KALMAN, Effective construction of linear state-variable model from input-output functions, Regelungstechnik , 12 (1966), pp. 545-8.
7. R.E.KALMAN, Mathematical description of linear dynamical systems, J.SIAM Control, 1 (1963), pp. 152-192.
8. R.E.KALMAN, Irreducible realizations and the degree of a rational matrix, J.SIAM Appl. Math., 13(1965), pp. 520-544.
9. R.E.KALMAN, On structure properties of linear constant multivariable systems, III IFAC Congress, London 1966, paper 6A.
10. Y.L.KUO,On the irreducible Jordan form realization and the degree of a rational matrix, IEEE Trans. on Circuit Theory, 17 (1970), pp. 322-332.
11. S.P.PANDA and C.T.CHEN, Irreducible Jordan form realization of a rational matrix, IEEE Trans. on Aut. Control, 14 (1969), pp. 66-9.
12. E.POLAK, An algorithm for reducing a linear-time-invariant differential system to state form, IEEE Trans. on Aut. Control, 11 (1966), pp. 577-9.
13. H.H.ROSENBROCK, Computation of minimal representation of rational transfer-function matrix, Proc. IEE, 115 (1968), pp. 325-7.
14. D.C.YOULA, The synthesis of linear dynamical systems from prescribed weighting pattern, J.SIAM Appl. Math., 14 (1966), pp. 527-549.

ON BANDPASS FUNCTIONS FROM LOWPASS POLES*

H. GORDON
Technical Consultant
Sarasota, Florida

G. W. ZOBRIST
University of South Florida
Tampa, Florida

ABSTRACT

A considerable amount of filter theory is based on the notion that for narrowband processes, the equivalent lowpass process can be used for purposes of analysis. In the narrowband process the complex conjugate pole clusters that must be added for realizability are neglected. In this paper a technique is introduced which minimizes the contribution of these complex conjugate poles.

ON BANDPASS FUNCTIONS FROM LOWPASS POLES

The standard lowpass to bandpass transformation distorts the arithmetic symmetry of the bandpass filter, however, if a translation of the poles is used for narrowband filters then the arithmetic symmetry is preserved to a large degree, even for large bandwidth[1]. Reference [2] discusses the effects of the standard lowpass to bandpass transformation on wideband arithmetically symmetric filters and presents a technique to approach arithmetic symmetry, whereby zeros of transmission are introduced at the origin for even ordered filters. A paper by Blinchikoff[3] gives a quantitative analysis of the lowpass to bandpass translation. Szentirmai[4,5] has presented general design techniques for arithmetically symmetric filters.

The equivalent bandpass process is defined to be the lowpass for ω under the transformation $p_i \to kp_i \pm j\omega_0$, where ω_0 is the band center and k scales the bandwidth. Clearly the function resulting from the transformation is rational and casual. The magnitude is given by

$$|T(s)|^2_{s=j\omega} = \frac{1}{Q(s)} \frac{1}{Q(-s)}\Big|_{s = j\omega}$$

Unfortunately, $1/Q(-s)$ has a contribution resulting from the third quadrant poles, and $1/Q(-s)$ has, by symmetry, the same contribution which results from the second quadrant poles of $\frac{1}{Q(s)}$.

---

*This paper represents a portion of an MSE thesis by the first author at the University of South Florida.

This paper presents a technique that will minimize this contribution, so that the poles of the bandpass magnitude function will approximate the poles of the lowpass magnitude function translated by $p_i \rightarrow kp_i + j\omega_0$.

Consider a single pole-pair of frequency $\omega_0 = 1$. This is equivalent to a one pole lowpass. Suppose a single transmission zero is allowed at some real frequency σ. Then the magnitude function is given by

$$|T(j\omega)|^2 = [\frac{1}{\alpha^2+(\omega-\beta)^2}] \; [\frac{\sigma^2+\omega^2}{\alpha^2+(\omega+\beta)^2}] \tag{1}$$

which is the translated lowpass multiplied by an error function. For small bandwidths $\beta \simeq \omega_0 = 1$, and

$$\alpha^2 << (1 + \omega)^2 \tag{2}$$

The pole-zero plot is given in Figure 1. Figure 2 shows how the arithmetic symmetry is distorted by the lowpass to bandpass translation. Thus the error function reduces to

$$|E(j\omega)| = \frac{(\sigma^2+\omega^2)^{1/2}}{1 + \omega}$$

Clearly the error is defined for any positive ω. Consequently, the derivative of the error is made to vanish for $\omega = \omega_0 = 1$ so that the error approximates a constant, which is acceptable. The values of σ are found by setting

$$\frac{d|E(j\omega)|}{d\omega} = \frac{\omega(1+\omega)}{(\sigma^2+\omega^2)} - 1 = 0$$

hence

$$\sigma^2 = 1 \; ; \; \sigma = \pm 1.$$

Either solution can be used depending upon the desired effect on group delay.

Now consider a double pole-pair under the same assumptions. This is equivalent to a two-pole lowpass. The magnitude function is given by

$$|T(j\omega)|^2 = [\frac{1}{\alpha^2 + (\omega-\beta_1)^2}] \; [\frac{1}{\alpha^2 + (\omega+\beta_2)^2}] \\ \times [\frac{\sigma^2 + \omega^2}{\alpha^2 + (\omega+\beta_1)^2}] \; [\frac{1}{\alpha^2 + (\omega+\beta_2)^2}]. \tag{3}$$

For small bandwidths

$$\beta_1, \beta_2 \simeq \omega_0 = 1 \tag{4}$$

Equations (2), (3), and (4) define the error function:

$$|E(j\omega)| = \frac{(\sigma^2 + \omega^2)^{1/2}}{(1 + \omega)^2}$$

As before, the error is defined for positive ω, and the derivative is made to vanish. The values of σ are found by setting

$$\frac{d|E(j\omega)|}{d\omega} = \frac{\omega\ (1 + \omega)}{\sigma^2 + \omega^2} - 2 = 0$$

hence

$$\sigma^2 = 0.$$

The transmission zero is at the origin.

The pole-zero plots and frequency response curves are similar to those presented in figure 1 and 2.

Since any higher order function can be described by a product expansion of terms like equations (1) and (3), we have a solution for higher order functions. This claim is justified by observing that the derivative of a product of a number of functions, all of which have a vanishing first derivative at some point, also vanishes at that point.

For a design example use the lowpass 3 pole cluster given in reference [1].

$p_1 = -.7263$

$p_{2,3} = -.6148 \pm j.9493$. For a desired bandwidth of 0.5 the translated poles are

$p_{1,2} = -.1816 \pm j1$

$p_{3,4} = -.1537 \pm j1.2373$

$p_{5,6} = -.1537 \pm j.7627$. These poles were obtained by using the translation $p_i = .25\ p_i \pm j1$. Table I compares the values of the magnitude function, for $\omega = .75$ and 1.25, before and after correction. Figure 3 is the pole-zero plot with the added zero's of transmission.

It should be further noted that for bandpass approximations which result in "symmetric" pole clusters, i.e. the highest and lowest frequency poles have the same (or very nearly) real part, the conclusions reached for the lowpass to bandpass translation still hold. Halpern [6] solved the flat time delay problem and obtained poles that meet the requirements set forth herein. He presents a comparison of translated lowpass flat time delay poles and those obtained with the direct approximation, and then presents graphs of the bandpass filters obtained from those roots.

In summary:

1. The order of the function is necessarily even.
2. If the order of the function is N, and if N/2 is even, there are N/4 transmission zeroes at the origin.
3. If the order of the function is N, and if N/2 is odd, there are (N-2)/4 transmission zeroes at the origin and one transmission zero at either $\sigma = +1$, or -1.

REFERENCES

(1) P. R. Geffe, "On the Approximation Problem for Bandpass Delay Lines", Proc. IRE, Vol. 50, No. 9, Sept. 1962.

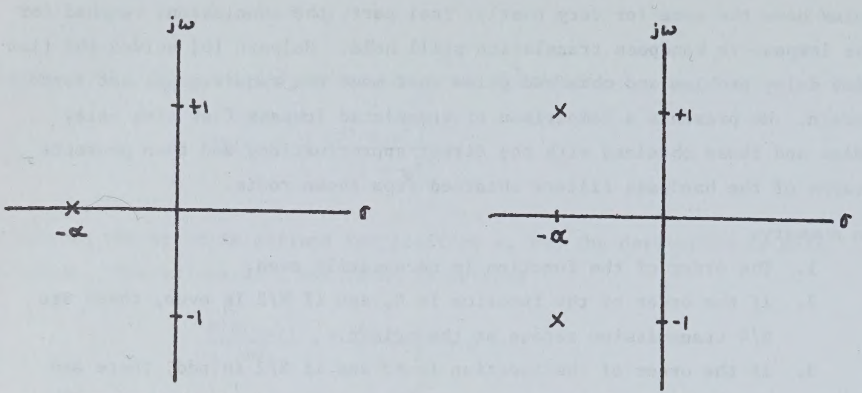
(2) W. M. Hummel, "Wide Band Linear Phase Band Pass Filters", Proc. 20th Electronic Components Conference, 1970.

(3) H. Blinchikiff, "A Note on Wide-Band Group Delay", to be published.

(4) G. Szentirmai, "The Design of Arithmetically Symmetrical Bandpass Filters", IEEE Trans. CT, Vol. CT-10, No. 3, Sept. 1963.

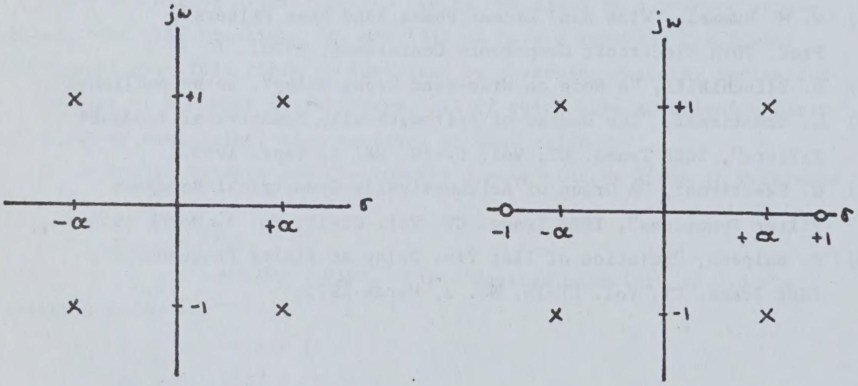
(5) G. Szentirmai, "A Group of Arithmetically Symmetrical Bandpass Filter Functions", IEEE Trans. CT, Vol. CT-11, No. 1, March 1964.

(6) P. Halpern, "Solution of Flat Time Delay at Finite Frequencies", IEEE Trans. CT, Vol. CT-18, No. 2, March 1971.


jω
+1
-1
σ
-α
jω
+1
-1
σ
-α

Lowpass pole-zero configuration

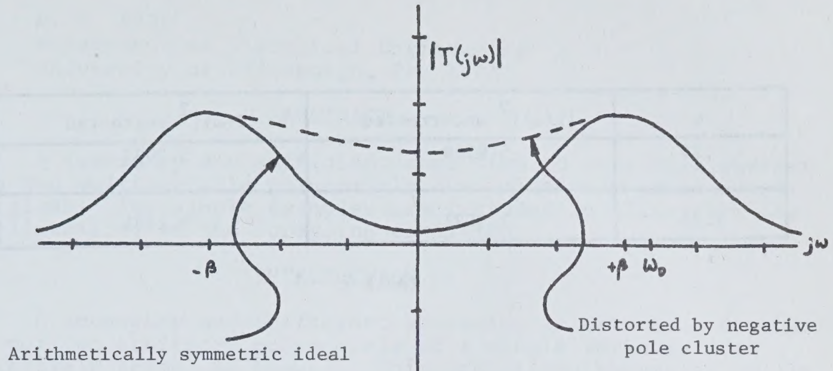
Equivalent Bandpass pole-zero configuration


jω
+1
-1
σ
-α
+α
jω
+1
-1
σ
-1
-α
+α
+1

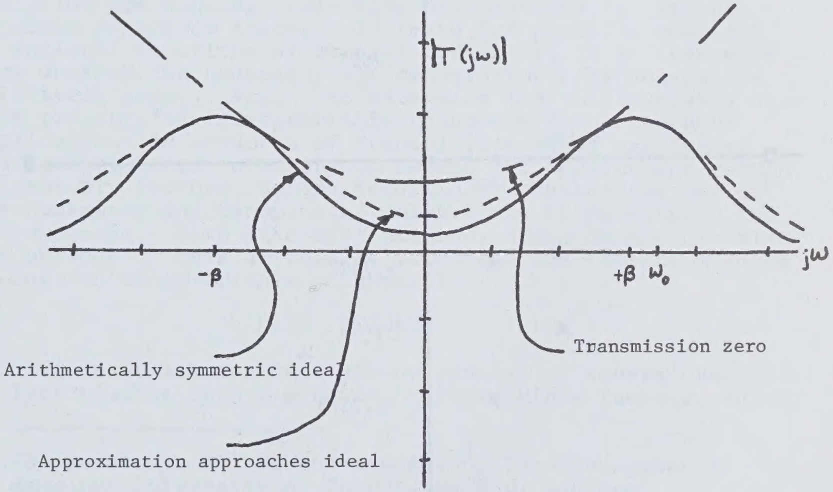
Bandpass pole-zero configuration with poles added in RHP for determining |T(s)|

Bandpass pole-zero configuration with zero at σ = ±1

Figure 1


|T(jω)|
jω
-β
+β
ω₀
Distorted by negative pole cluster
Arithmetically symmetric ideal

Magnitude curves without zero correction

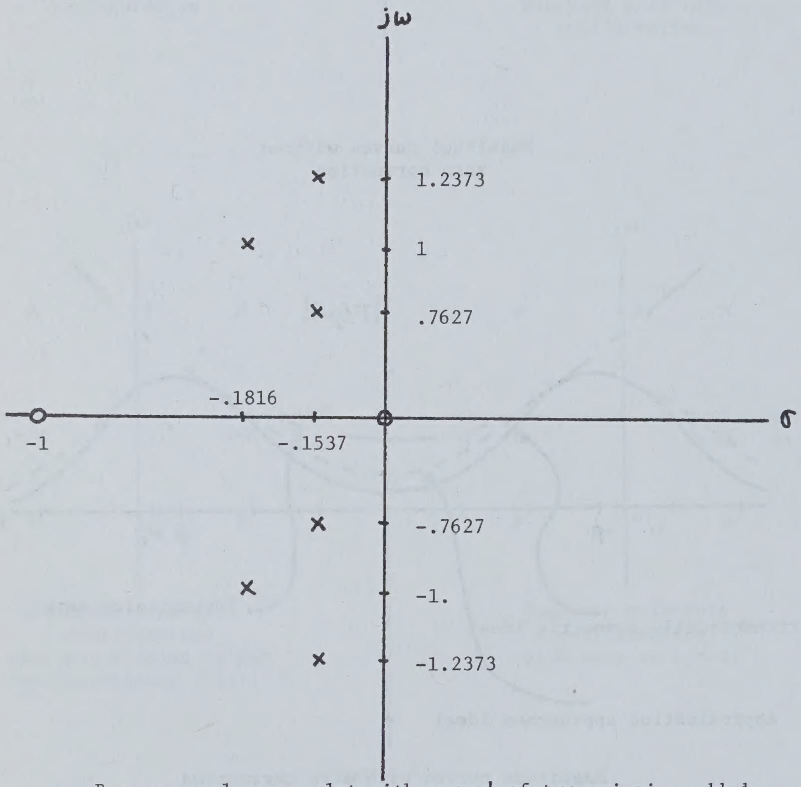

|T(jω)|
jω
-β
+β
ω₀
Transmission zero
Arithmetically symmetric ideal
Approximation approaches ideal

Magnitude curves with zero correction

Figure 2

ω	$\|T(\omega)\|^2$ uncorrected	$\|T(\omega)\|^2$ corrected
.75	5.9269	5.2092
1.25	1.3071	5.2335

TABLE I


jω
1.2373
1
.7627
-.1816
-1
-.1537
σ
-.7627
-1.
-1.2373

Banspass pole-zero plot with zeros' of transmission added

Figure 3

A CRITERION TO DETERMINE IF TWO MULTIVARIABLE POLYNOMIALS ARE RELATIVELY PRIME

N. K. BOSE*
Department of Electrical Engineering
University of Pittsburgh, Pa. 15213

ABSTRACT

A necessary and sufficient condition to determine whether any two multivariable polynomials are relatively prime or not is given. Two simple examples are included to illustrate the applicability of the foregoing condition

INTRODUCTION

A necessary and sufficient condition to determine whether or not two arbitrary polynomials of a single variable are relatively prime, is known[1]. This condition, expressed solely and directly in terms of the polynomial coefficients requires the evaluation of the determinant of a matrix of smaller order than the matrix required in the evaluation of the resultant[2]. The result, independently, arrived at by several authors[3] is written below in the form of a lemma to facilitate reading.

Lemma If f(p) and g(p) are two monic polynomials of a single variable p, then a necessary and sufficient condition for the preceding polynomials to have a common factor is that det f(B) = 0 or det g(A) = 0, where A and B are, respectively, the companion matrices associated with the polynomials f(p) and g(p).

At present, polynomials and rational functions of several variables are finding increasing applications in various problems in system theory. In tests for positive reality[4] of rational functions of several variables, it is necessary to know whether the numerator and denominator polynomials are relatively prime. Again, to determine the multivariable reactance property[4.5] (multivariable reactance functions have applications in networks of transmission lines and lumped reactance), it is necessary to cancel all nonconstant common polynomial factors, before determining whether the sum of the numerator and denominator polynomials is Hurwitz in the narrow sense. With like applications, among others, in mind the purpose of this article is to extend the foregoing lemma to any two multivariable polynomials.

MAIN RESULTS

It is well known that any polynomial of several variables is factorizable into a product of irreducible factors, which are uniquely determined, regardless of their order and constant factors.[6] Also, a given polynomial $H(p_1,p_2,\ldots p_n)$ of n variables can always be written in the form,

---

* The author is at present a visiting faculty member at the American University of Beirut, Beirut, Lebanon

$$H(p_1, p_2,\ldots p_n) = f_{0i}p_i^{\,k} + f_{1i}p_i^{\,k-1} + \ldots + f_{ki} \qquad 1 \leqslant i \leqslant n \quad (1)$$

where k is any positive integer or zero, and f_{0i},lllf_{ki} are each polynomials in the (n-1) variables $p_1, p_2,\ldots p_{i-1}, p_{i+1}, \ldots p_n$ and f_{0i} is not identically zero. Given two polynomials $H(p_1, p_2,\ldots p_n)$ and $R(p_1, p_2,\ldots p_m)$ it is always possible to write these in the form,

$$\begin{aligned} H(p_1, p_2,\ldots p_n) &= f_{0i}(p_i^{\,k} + a_{1i}{}^{k-1} + \ldots + a_{ki}) \\ &= f_{0i}H_i(p_1, p_2,\ldots p_n), \quad \text{and} \qquad (2) \\ R(p_1, p_2,\ldots p_m) &= g_{0i}(p_i^{\,\ell} + b_{1i}p_i^{\,\ell-1} + \ldots + b_{\ell i}) \\ &= g_{0i}R_i(p_1, p_2,\ldots p_m) \end{aligned}$$

where $1 \leqslant i \leqslant m$, and also $1 \leqslant i \leqslant n$.

Furthermore, $a_{1i} = \dfrac{f_{1i}}{f_{0i}}$, $a_{2i} = \dfrac{f_{2i}}{f_{0i}}$, ... etc.

and the polynomial coefficients in the expression for $R(p_1, p_2,\ldots p_m)$ are analogously defined. It is noted that $H_i(p_1, p_2,\ldots p_n)$ and $R_i(p_1, p_2,\ldots p_m)$ are monic polynomials in the variable p_i, with not necessarily polynomial (but, in general, reational) coefficients in the remaining variables. In some cases it might be possible to rewrite the expressions for $H(p_1, p_2,\ldots p_n)$ and $R(p_1, p_2,\ldots p_m)$ in the following forms,

$$\begin{aligned} H(p_1, p_2,\ldots p_n) &= f_{0i}H_i(p_1, p_2,\ldots p_n) = f_{0i1}f_{0i2}H_i(p_1,p_2,\ldots p_n) \\ &= f_{0i1}H_{i1}(p_1, p_2,\ldots p_n), \quad \text{and} \qquad (3) \\ R(p_1, p_2,\ldots p_m) &= g_{0i}R_i(p_1, p_2,\ldots p_m) = g_{0i1}g_{0i2}R_i(p_1,p_2,\ldots p_m) \\ &= g_{0i1}R_{i1}(p_1, p_2,\ldots p_m) \end{aligned}$$

where $H_{i1}(p_1, p_2,\ldots p_n)$ and $R_{i1}(p_1, p_2,\ldots p_m)$ are not necessarily monic polynomials in p_i, but each have polynomial coefficients in the remaining variables.

Associate companion matrices A_i and B_i, whose elements are functions of all variables involved except p_i, such that

$$\begin{aligned} \det(p_iI_k - A_i) &= H_i(p_1, p_2,\ldots p_n) = p_i^{\,k} + a_{1i}p_i^{\,k-1} + \ldots + a_{ki}, \text{and} \\ \det(p_iI_\ell - B_i) &= R_i(p_1, p_2,\ldots p_m) = p_i^{\,\ell} + b_{1i}p_i^{\,\ell-1} + \ldots + b_{\ell i} \qquad (4) \end{aligned}$$

with $1 \leqslant i \leqslant m$, and $1 \leqslant i \leqslant n$. Furthermore I_k and I_ℓ are identity matrices of orders k and ℓ, respectively. Then the following theorem follows.

Theorem: A necessary and sufficient condition for two polynomials, $H(p_1, p_2, \ldots p_n)$ and $R(p_1, p_2, \ldots p_m)$ to have a common factor is that for any one i, where $1 \leqslant i \leqslant n$, and also $1 \leqslant i \leqslant m$,

(a) $\det(B_i{}^k + a_{1i}{}^{k-1} + \ldots + a_{ki}I_\ell) = 0$, or

(b) $\det(A_i{}^\ell + b_{1i}A_i{}^{\ell-1} + \ldots + b_{\ell i}I_k) = 0$, or

(c) $f_{0i1} = g_{0i1}$, or

(d) f_{0i1} and g_{0i1}, which in the most general case are polynomials with one fewer variable than $H(p_1, p_2, \ldots p_n)$ and $R(p_1, p_2, \ldots p_m)$, respectively, and on which tests similar to that in (a) or (b) have to be repeated, have a common factor.

The proof of the above theorem is simple and is omitted for brevity, but can be obtained along lines, to a great degree parallel to the single variable case. It is advisable before repeating the cycle of operation in (a) or (b), to check for the presence of a common factor in f_{0i1} and g_{0i1}, in (d) by inspection. The examples below illustrate the applicability of the above theorem.

EXAMPLES

Example 1: It is required to determine whether the polynomials,

$$H(p_1, p_2, p_3) = p_1{}^2p_2p_3{}^2 + p_3{}^2p_1 + 2p_1{}^2p_2 + 2p_1 \quad \text{and}$$

$$R(p_1, p_2) = p_1{}^3p_2{}^3 + p_1{}^2p_2{}^2 + 4p_1p_2 + 4,$$ are relatively prime or not.

Here n = 3, m = 2. For i = 1, the above two polynomials can be written in the form;

$$H(p_1, p_2, p_3) = (p_2p_3{}^2 + 2p_2)p_1{}^2 + (p_3{}^2 + 2)p_1$$

$$R(p_1, p_2) = p_2{}^3p_1{}^3 + p_2{}^2p_1{}^2 + 4p_2p_1 + 4.$$

Consequently, $f_{01} = p_2(p_3{}^2 + 2)$, $g_{01} = p_2{}^3$, $k = 2$, $\ell = 3$.

$$a_{11} = \frac{1}{p_2}, \quad b_{11} = \frac{1}{p_2}, \quad b_{21} = \frac{4}{p_2{}^2}, \quad b_{31} = \frac{4}{p_2{}^3}$$

$$A_1 = \begin{bmatrix} \dfrac{-1}{p_2}, & 0 \\ 1 & 0 \end{bmatrix}, \text{ and}$$

$$\det(A_1{}^3 + b_{11}A_1{}^2 + b_{21}A_1 + 4I_2) = \det\begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} = 0$$

Therefore $H(p_1, p_2, p_3)$ and $R(p_1, p_2)$ are not relatively prime.

Example 2: It is required to determine if

$H(p_1, p_2) = p_1^3 p_2 + 3p_1^2 + p_1 p_2 + 3$ and

$R(p_1, p_2) = p_1^3 p_2 + 2p_1^2 + p_1 p_2 + 2$
are relatively prime or not.

Here $n = m = 2$. For $i = 2$, the above two polynomials take the form:

$$H(p_1, p_2) = (p_1^3 + p_1)p_2 + (3p_1^2 + 3) = (p_1^3 + p_1)(p_2 + \frac{3}{p_1})$$

$$= (p_1^2 + 1)p_1(p_2 + \frac{3}{p_1}) = (p_1^2 + 1)(p_1 p_2 + 3)$$

and

$$R(p_1, p_2) = (p_1^3 + p_1)p_2 + (2p_1^2 + 2) = (p_1^3 + p_1)(p_2 + \frac{2}{p_1})$$

$$= (p_1^2 + 1)p_1(p_2 + \frac{2}{p_1}) = (p_1^2 + 1)(p_1 p_2 + 2)$$

It is seen that, $f_{0i1} = (p_1^2 + 1) = g_{0i1}$.

Therefore, the two polynomials are not relatively prime.

CONCLUSION

A necessary and sufficient condition to determine whether any two polynomials of several variables are relatively prime or not has been given. No problems arise if singularities of a second kind due to Weirstrass are present in the two polynomials under consideration. For example, application of the theorem to the two polynomials,

$H(p_1, p_2) = p_2^2 + 2p_1 p_2 + 1$, and

$R(p_1, p_2) = (p_1 + 1) p_2^2 + p_1(p_1 + 2) p_2 + p_1(p_1 + 1)$, which have a common zero at $p_1 = 0$, $p_2 = \pm j$, but no common factors, indeed reveals the fact that $H(p_1, p_2)$ and $R(p_1, p_2)$ are relatively prime. The ease of application of the theorem is enhanced by the fact that the results of the theorem are independent of i, so long as $1 \leq i \leq n$ and $1 \leq i \leq m$.

REFERENCES

1. W.G. Vogt and N.K. Bose, "A Method to Determine Whether Two Polynomials are Relatively Prime," IEEE Trans. on Automatic Control, June 1970, pp. 379-380.

2. W.S. Burnside and A.W. Panton, "The Theory of Equations," vol. 2, New York, Dover, 1960.

3. S. Barnett, "Comments on 'A Method to Determine Whether Two Polynomials are Relatively Prime,'" IEEE Trans. on Automatic Control, Feb. 1971, p. 113.

4. T. Koga, "Synthesis of Finite Passive n-Ports with Prescribed Real Matrices of Several Variables," IEEE

4. Transaction on Ckt. Theory, March, 1968, pp. 2-24.

5. H.G. Ansell, "On Certain Two-Variable Generations of Circuit Theory, with Applications to Networks of Transmission Lines and Lumped Reactances," IEEE Trans. on Ckt. Theory, June 1964, pp. 214-224.

6. G.A. Bliss, "Algebraic Functions," Am. Math. Sec., New York N.Y.: 1933.

AN OPTIMAL ITERATIVE METHOD FOR THE COMPUTATION OF MATRIX RANK

NARESH N. GUPTA
Wright State University
Dayton, Ohio 45431

Abstract: An optimal iterative method for computing rank of any (real or complex) matrix A is developed. The method makes use of an upper bound on the maximum eigenvalue of AA* (or A*A). As an illustration we compute the rank of 15 x 15 Hilbert matrix.

The problem of determining [1,2,3,4] the order of a linear differential system from input-output data reduces to finding the rank of certain matrix or matrices. In [8,9] we gave a method for computing generalized inverse A^+ for any matrix A and also showed how the sequence of approximations to AA^+ (or A^+A) could be used to compute the rank of matrix A. Below we give a Theorem which gives another way of implementing the computation of matrix rank and is more accurate than the standard Gaussian elimination technique or the algorithm given by Theorems 2 and 3 in [8,9].

Notation: Let A denote an n x m nonzero complex matrix, A^* its conjugate transpose $r(\leq n)$ = rank A.

Let $\lambda_1(AA^*) \geqslant \lambda_2(AA^*) \geqslant \ldots \geqslant \lambda_n(AA^*)$ be the eigenvalues of AA^*. From rank A = r it follows that $\lambda_r(AA^*) > 0$ and $\lambda_i(AA^*) = 0$ for i = r + 1, ..., n. We will use the spectral norm [5] $||A|| = \lambda_1^{1/2}(AA^*)$.

Theorem 1:

The trace of $[I - \alpha AA^*]^{2k}$ is a monotone decreasing function of k, k = 1,2,... converging to the nullity of matrix A with $0<\alpha< 2/\lambda_1(AA^*)$ where $\lambda_1(AA^*)$ is the maximum eigenvalue of AA^* (or A^*A).

Proof:

$$\text{let } T_k = \text{Trace } [I - \alpha AA^*]^{2k} = n - r + \sum_{i=1}^{r} (1 - \alpha\lambda_i)^{2k}$$

where λ_i's are the nonzero eigenvalues of AA^* (or A^*A). Therefore $T_{k+1} \leqslant T_k$.

We call α_0 optimal if it minimizes $||I - \alpha AA^*||$. The function $G(\alpha) = ||I - \alpha AA^*||$ is convex and $G(0) = G(2/\lambda_1 (AA^*)) = 1$. It can be shown [6] that $G(\alpha)$ has a unique minimum in the interval $0<\alpha< 2/\lambda_1 (AA^*)$.

Theorem 2: [10] The optimal α is

$$\alpha_0 = \frac{2}{\lambda_1(AA^*) + \lambda_r(AA^*)}$$

Proof:

The minimizing α_0 must satisfy $1 - \alpha\lambda_r(AA^*) = -(1 - \alpha\lambda_1(AA^*))$, i.e., the interval $[\lambda_r(AA^*), \lambda_1(AA^*)]$ is mapped onto an interval symmetric around the origin.

Remark 1: In using the results of this paper, it is not necessary to compute $\lambda_1(AA^*)$--the maximum eigenvalue of AA^* (or A^*A). Writing $AA^* = (C_{ij})$ $(i,j = 1,2,\ldots,n)$ we conclude from the Gershgorin theorem that $\lambda_1(AA^*) < \max_i \left[\sum_{j=1}^{n} |C_{ij}|\right]$ and condition on α can be replaced by $0<\alpha<2/\max_i \left[\sum_{j=1}^{n} |C_{ij}|\right]$.

Remark 2: Note that ratio of Trace $[I - \alpha AA^*]^{2^k}$ to Trace $[I - \alpha AA^*]^{2^k + 1}$ tends to unity for large values of k when the matrix A is of less than full rank and Trace $[I - AA^*]^{2^k}$ tends to zero for large values of k when the matrix A is of full rank.

Using Theorem 1 to determine the rank of any matrix A, we need to compute the sequence of traces for $[I - \alpha AA^*]^{2^k}$ or $[I - \alpha A^*A]^{2^k}$ whichever is of lower order.

Example: The computation of matrix rank is demonstrated below. The computations were carried out on an IBM 360-50, using double precision arithmetic. The matrix is

$$A = (a_{ij}) = \frac{1}{(i + j - 1)} \quad ; \ i,j = 1,2,\ldots, 15.$$

The results of computation are summarized in a table below. The numbers in the parenthesis should be read as exponent, e.g., 0.1000(-01) is 0.1×10^{-1}.

TABLE

Computation of Rank of 15x15 Hilbert Matrix

i	1	2
α_i	$\alpha_1 = 0.233000$	$\alpha_2 = 0.300000$
k	Trace $(I - \alpha AA^*)^{2^k}$	Trace $(I - \alpha AA^*)^{2^k}$
1	0.138394806(+02)	0.137948729(+02)
10	0.121944511(+02)	0.121142429(+02)
20	0.109107508(+02)	0.108865815(+02)
30	0.969139227(+01)	0.962166012(+01)
40	0.838421560(+01)	0.829130222(+01)
50	0.645377305(+01)	0.635965817(+01)
52	0.465539503(+01)	0.457439089(+01)
53	0.327554911(+01)	0.316194893(+01)
54	0.184992651(+01)	0.167894243(+01)
55	0.806617789(+00)	0.580760313(+00)
56	0.256741907(+00)	0.929219807(-01)
57	0.465708245(-01)	0.326037505(-02)
58	0.207995428(-02)	0.545890229(-05)
59	0.432338460(-05)	0.211841682(-10)
60	0.186916507(-10)	

Since few eigenvalues [7] of A are $<10^{-13}$ the convergence is slow. About 60 iterations are needed to converge to the nullity of matrix A.

The round-off errors in implementing Theorem 1 are smaller than those in implementing the standard Gaussian elimination technique. For example, Gaussian elimination technique gave inaccurate rank for Hilbert matrix of order $\geqslant 10$. Therefore, this method is more reliable than standard Gaussian elimination technique. However, this method takes relatively more computation time, but the Hermitian property of $[I - \alpha AA^*]^{2^k}$

can be exploited to further reduce the computation time for matrix multiplication.

The round-off errors in implementing Theorem 1 are also smaller than those in implementing the algorithm given by Theorems 2 and 3 in [8,9]. For example, in the case of 15x15 Hilbert matrix no convergence was obtained by using Theorems 2 and 3 in [8,9].

The following is an amusing example of the sensitivity to the choice of α of Theorems 1 and 2 in [8,9].

Let A be any nxm partial isometry of rank r. Then we know that $A^+ = A^*$ and $\lambda_{max} = \lambda_{min} = 1$. Hence $\alpha = 1$ gives convergence in no iterations (i.e., $S_0 = A^+$) while the trace of $(I - \alpha AA^*)^{2^k} = r(1-\alpha)^{2^k} + n-r$, shows the convergence for any $0<\alpha<2$.

References

1. Naresh N. Gupta, "On the order discrimination of linear systems from noisy observations," to appear in the Proceedings of the IEEE Decision and Control Conference, Gainsville, Florida, Dec. 1971.
2. K.J. Astrom and T. Bohlin, "Numberical identification of linear dynamic system from normal operating records," paper IFAC symposium on Self-Adaptive control Systems, Teddington, England, pp. 41-51, Sept., 1966.
3. B.L. Ho and R.E. Kalman, "Effective construction of linear state-variable models from input-output functions," Regelungstechnik, vol. 12, pp. 545-548, 1965.
4. R.C.K. Lee, "Optimal estimation, identification and control," M.I.T. Press, 1964.
5. L. Collatz, "Functional analysis and numberical mathematics," Academic Press, New York, 1966.

6. M. Altman, "An optimum cubically convergent iterative method of inverting a linear bounded operator in Hilbert space," Pacific J. Math., vol. 10, pp. 1107-1113, 1960.

7. H.E. Fettis and J.C. Caslin, "Eigenvalues and eigenvectors of Hilbert matrices of order 3 through 10, " Math. Comp., Vol. 21, no. 99, pp. 431-441, 1967.

8. Naresh N. Gupta, "On the convergence of an iterative method for the computation of generalized inverse and associated projections, International Journal of Systems Science, 1971 Vol. 2, No. 1, pp. 67-75.

9. Naresh N. Gupta, "An iterative method for computation of generalized inverse and matrix rank", I.E.E.E. Transactions, on Systems Man and Cybernetics Vol. SMC - 1, No. 1, Jan. 1971, pp. 89-90.

10. Naresh N. Gupta, "An optimum hyperpower method for computing generalized inverses and associated projections," to appear in Int. J. Systems Sci. 1972.

A NOTE ON THE GENERALIZED HAMILTON-JACOBI EQUATION FOR DIFFERENTIAL GAMES WITH INFORMATION TIME LAG*

MICHAEL D. CILETTI
F. J. Seiler Research Laboratory (AFSC)
USAF Academy, Colorado 80840

ABSTRACT

The generalized Hamilton-Jacobi equation over the τ-t data reference plane in differential games with information time lag will be reduced to a partial differential equation along the line $\tau = t-\sigma$, for fixed σ. Application of these results to a linear regulator problem will show that the generalized Riccati equation in the variables t and τ can be solved along the line $\tau = t-\sigma$. This makes possible the design of feedback compensators which account for the "age" of the state variable data.

1. INTRODUCTION

Some of the pertinent highlights of the theory of differential games with information time lag (DGWITL) can be described briefly as follows. A detailed, rigorous and more general development can be found in reference [1].

Given that the pursuer (P) and the evader (E) control the dynamic systems described by:

$$\dot{x} = f(x,\tilde{u},t) \tag{1}$$

$$\dot{y} = g(y,\tilde{v},t) \tag{2}$$

where $\tilde{u}$ and $\tilde{v}$ are instantaneously specified by strategies u and v which are members of an appropriate class of admissible strategies, we associate with P and E's behavior in controlling the system the following payoff made by P to E upon completion of play:

$$J(x,t,y,t,u,v) = \int_t^{t_p} L_1(x,\tilde{u},\alpha)\, d\alpha$$

$$\int_t^{t_p} L_2(y,\tilde{v},\alpha)\, d\alpha + \tilde{W}(x(t_p),t_p,y(t_p),t_p), \tag{3}$$

with t_p fixed. The redundant appearance of time arguments in the above expressions is allowed so that control laws obtained later will reflect the "age" of the state variable in a problem with information time lag. In particular, we are concerned with a game beginning at some time t in which

---

*The views expressed herein are those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

E has the following data observation tuple: (x,τ,y,t), where $\tau \leq t$, and (x,τ) is a phase of (1) and (y,t) is a phase of (2). Given that E knows at time t his position, y, and that at time τ P's position was x, is there an optimal way for E to play the game? If E is provided with a continuous record of P's motion, i.e., at each time t E receives $x(t-\sigma)$, σ fixed, can we construct an optimal feedback controller for E and an optimal exploitation strategy for P of the form: $v^\circ(x,t-\sigma,y,t)$, $u^\circ(x,t-\sigma,y,t)$, respectively?

2. GENERALIZED HAMILTON-JACOBI EQUATION

We have shown in [1] that the important quantity for E to consider at each time t is that portion of J which remains after we remove those terms that can be computed with certainty by E at time t. Hence we define the remaining payoff:

$$J_r(x,\tau,y,t,u,v) = \int_\tau^{t_p} L_1(x,u,\alpha)\,d\alpha + \int_t^{t_p} L_2(y,v,\alpha)\,d\alpha$$

$$+ \tilde{W}(x,t_p,y,t_p), \tag{4}$$

where $\tau = t-\sigma$, and $\tilde{W}(\cdot)$ is evaluated at time t_p on the motions of (1) and (2) which emanate from (x,τ) and (y,t). We then form the potential payoff function, J_p, as:

$$J_p(x,\tau,y,t,u,v) = \int_\tau^{t_p-\sigma} L_1(x,u,\alpha)\,d\alpha + \int_t^{t_p} L_2(y,v,\alpha)\,d\alpha$$

$$+ W(x,t_p-\sigma,y,t_p) \tag{5}$$

where $W(\cdot)$, the potential terminal payoff, is the solution to the following optimal control problem:

$$W(\cdot) = \min_u \int_{t_p-\sigma}^{t_p} L_1(x,u,\alpha)\,d\alpha + \tilde{W}(x,t_p,y,t_p)$$

subject to $\dot{x} = f(x,u,t)$ with $(x,t_p-\sigma)$ as initial data. We are able to show that if J_p has a saddlepoint w.r.t. admissible u and v, then the following generalized Hamilton-Jacobi equation (GHJE) must be satisfied where the saddlepoint of the potential value, $V^\circ(x,\tau,y,t)$, is smooth:

$$\frac{\partial V^\circ}{\partial \tau} + \frac{\partial V^\circ}{\partial t} + H^\circ(x,\tau,y,t,V^\circ_x,V^\circ_y) = 0 \tag{6}$$

subject to: $V^\circ(x,t_p-\sigma,y,t_p) = W(x,t_p-\sigma,y,t_p)$, where:

$$H^\circ(x,\tau,y,t,V_x^\circ,V_y^\circ) = H(x,\tau,y,t,V_x^\circ,V_y^\circ,k_1,(x,\tau,y,t,V_x^\circ,V_y^\circ),k_2(x,\tau,y,t,V_x^\circ,V_y^\circ)),$$

$$\min_\mu \max_\beta H(x,\tau,y,t,\mu,\beta,\lambda_x,\lambda_y) = \max_\beta \min_\mu H(x,\tau,y,t,\mu,\beta,\lambda_x,\lambda_y) =$$

$$= H(x,\tau,y,t,k_1\ (x,\tau,y,t,\lambda_x,\lambda_y),k_2(x,\tau,y,t,\lambda_x,\lambda_y),\lambda_x,\lambda_y)$$

and:

$$H(\cdot) = <\lambda_x,f(x,\mu,\tau)> + <\lambda_y,g(y,\beta,t)> + L_1(x,\mu,\tau) + L_2(y,\beta,t)$$

with $\mu\ \varepsilon\ A_u \subset R^q,\ \beta\ \varepsilon\ A_v \subset R^m, \lambda_x,\lambda_y\ \varepsilon\ R^n$.

The generalized Hamilton-Jacobi equation is defined on the playing space for P and E and over the τ-t "data reference plane." The novel feature of this equation is the presence of τ, the time corresponding to the delayed state variable, x. The equation also specifies the optimal state feedback control law for the evader which implements the data (x,τ,y,t); i.e., the control law accounts for the age of data.

In general there is no way to solve the GHJE globally. However, we will now show that one can easily reduce the GHJE in τ and t to a partial differential equation in t. This equation describes the behavior of V° along the line $\tau = t-\sigma$ for fixed σ and is relevant to actual play in which a fixed data lag is present.

3. POTENTIAL VALUE ALONG $\tau = t-\sigma$

To examine $V^\circ(\cdot)$ along the locus for a fixed data lag in the data reference plane, let $\tau = t-\sigma$, with σ fixed, and let $\hat{V}^\circ(x,y,t;\sigma)$ denote $V^\circ(x,\tau,y,t)$ along $\tau = t-\sigma$, i.e.,

$$\hat{V}^\circ(x,y,t;\sigma) = V^\circ(x,\tau,y,t)\Big|_{\tau=t-\sigma}$$

Likewise, let

$$\hat{H}^\circ(x,y,t,V_x^\circ,V_y^\circ;\sigma) = H^\circ(x,\tau,y,t,V_x^\circ,V_y^\circ)\Big|_{\tau=t-\sigma}$$

Then:

$$\frac{\partial\hat{V}^\circ}{\partial t} = \frac{\partial V^\circ}{\partial\tau}\frac{d\tau}{dt} + \frac{\partial V^\circ}{\partial t} \qquad = \frac{\partial V^\circ}{\partial\tau} + \frac{\partial V^\circ}{\partial t}$$

$$\frac{\partial\hat{V}^\circ}{\partial x} = \frac{\partial V^\circ(x,\tau,y,t)}{\partial x}\Big|_{\tau=t-\sigma} \qquad \frac{\partial\hat{V}^\circ}{\partial y} = \frac{\partial V^\circ(x,\tau,y,t)}{\partial y}\Big|_{\tau=t-\sigma}$$

and so (6) becomes

$$\frac{\partial \hat{V}^\circ}{\partial t}(x,y,t;\sigma) + \hat{H}^\circ(x,y,t,\hat{V}^\circ_x(x,y,t;\sigma),\hat{V}^\circ_y(x,y,t;\sigma);\sigma) = 0 \qquad (7)$$

Subject to:

$$\hat{V}^\circ(x,y,t_p;\sigma) = W(x,t_p-\sigma,y,t_p)$$

Whereas (6) describes $V^\circ(\cdot)$ over the region of τ-t plane below $\tau=t$, (7) describes $V^\circ(\cdot)$ on a line corresponding to play with fixed σ. More importantly, (7) is a partial differential equation on x,y,t and the gradient of $\hat{V}^\circ(\cdot)$, with parameter σ. Thus (7) is amenable to the derivation of characteristic equations and the establishment of an indirect method for solving DGWITL.

4. LINEAR REGULATOR REVISITED

We now present the linear regulator as an example to illustrate the application of the theory of DGWITL.

The systems are described by:

$$\dot{x} = A_p(t)x(t) + B_p(t)u(t) \; ; \; x(\tau_o) = x_{o\sigma}$$

$$\dot{y} = A_e(t)y(t) + B_e(t)v(t) \; ; \; y(t_o) = y_o$$

with $u(t) \in A_u \subset R^q$, $v(t) \in A_v \subset R^m$. The remaining payoff function is

$$J_r(x_{o\sigma},\tau_o,y_o,t_o,u,v) = ||x(T) - y(T)||^2_F +$$

$$\int_{\tau_o-\sigma}^{T} ||u||^2_{R_p(\alpha)}d\alpha - \int_{t_o}^{T} ||v||^2_{R_e(\alpha)}d\alpha$$

where $F>0, R_p(\cdot)>0, R_e(\cdot)>0$, and $||u||^2_{R_p(\alpha)}$ denotes $u'(\alpha)R_p(\alpha)u(\alpha)$.

For a fixed x,y we examine motions of (1) over $[T-\sigma,T]$ emanating from $(x,T-\sigma)$ and define the potential terminal payoff:

$$W(x,T-\sigma,y,T) = \min_u ||x(T)-y||^2_F + \int_{T-\sigma}^{T} ||u||^2_{R_p(\alpha)}d\alpha$$

Function space or other methods are easy to apply to this problem, and we simply provide here the result:

$$W(x,T-\sigma,y,T) = [\Phi_p(T,T-\sigma)x-y]'[F^{-1} + SR_p^{-1}S^*]^{-1}[\Phi_p(T,T-\sigma)x-y]$$

where

$$SR_p^{-1}S^* \triangleq \int_{T-\sigma}^{T} \Phi_p(T,\alpha)B_p(\alpha)\, R_p^{-1}(\alpha)B_p'(\alpha)\Phi_p'(T,\alpha)\,d\alpha$$

and $\Phi_p(\cdot)$ is the transition matrix for $A_p(\cdot)$. The optimal terminating control, $\bar{u}$, is given by

$$\bar{u}(\alpha) = -R_p^{-1}(\alpha)B_p'(\alpha)\Phi_p'(T,\alpha)[F^{-1} + SR_p^{-1}S^*]^{-1}[\Phi_p(T,T-\sigma)x-y], \alpha\varepsilon[T-\sigma,T]$$

It follows that the potential payoff is simply:

$$J_p(x_{o\sigma},\tau_o,y_o,t_o,u,v) = [\Phi_p(T,T-\sigma)x-y]'[F^{-1} + SR_p^{-1}S^*]^{-1}[\Phi_p(T,T-\sigma)x-y]$$

$$+ \int_{t_o-\sigma}^{T-\sigma} ||u||^2_{R_p(\alpha)}\,d\alpha - \int_{t_o}^{T} ||v||^2_{R_e(\alpha)}\,d\alpha$$

where $x = x(T-\sigma)$ and $y = y(T)$. To apply the theory for DGWITL to this problem we form the generalized pre-Hamiltonian:

$$H(x,\tau,y,t,\mu,\beta) = \langle A_p(\tau)x + B_p(\tau)\mu, \lambda_x\rangle +$$

$$\langle A_e(t)y + B_e(t)\beta, \lambda_y\rangle + \langle\mu, R_p(\tau)\mu\rangle - \langle\beta, R_e(t)\beta\rangle$$

The usual min max operation leads to:

$$k_1(x,\tau,y,t,\lambda_x,\lambda_y) = -\frac{1}{2}R_p^{-1}(\tau)B_p'(\tau)\lambda_x$$

$$k_2(x,\tau,y,t,\lambda_x,\lambda_y) = \frac{1}{2}R_e^{-1}(t)B_e'(t)\lambda_y$$

and the following generalized Hamiltonian

$$H^\circ(x,\tau,y,t,\lambda_x,\lambda_y) = x'A_p'(\tau)\lambda_x - \frac{1}{4}\lambda_x'B_p(\tau)R_p^{-1}(\tau)B_p'(\tau)\lambda_x$$

$$+ \frac{1}{4}\lambda_y'B_e(t)R_e^{-1}(t)B_e'(t)\lambda_y + y'A_e'(t)\lambda_y$$

The generalized Hamilton-Jacobi equation satisfied by the potential value is:

$$V^\circ_\tau + V^\circ_t + x'A_p'(\tau)V^\circ_x + y'A_e'(t)V^\circ_y$$

$$-\frac{1}{4}V^{\circ\prime}_x B_p(\tau)R_p^{-1}(\tau)B_p'(\tau)V^\circ_x + \frac{1}{4}V^{\circ\prime}_y B_e(t)R_e^{-1}(t)B_e'(t)V^\circ_y = 0$$

with boundary condition:

$$V^\circ(x,\tau,y,t)\Big|_{\substack{\tau=T-\sigma\\ t=T}} = [\Phi_p(T,T-\sigma)x - y]'[F^{-1} + SR_p^{-1}S^*]^{-1}[\Phi_p(T,T-\sigma)x - y]$$

$$= \begin{bmatrix} x \\ y \end{bmatrix}' \left[\begin{array}{c} \Phi_p'(T,T-\sigma) \\ \hline -I \end{array}\right] [F^{-1} + SR_p^{-1}S^*]^{-1} [\, \Phi_p(T,T-\sigma) \mid -I\,] \begin{bmatrix} x \\ y \end{bmatrix}$$

where I denotes the identity matrix for R^n. We point out here that letting $\sigma \to 0$ and $\tau \to t$ produces the well known results for DG without time lag. Following the procedures for linear optimal control and DG we next assume that

$$V^\circ(x,\tau,y,t) = \begin{bmatrix} x \\ y \end{bmatrix}' P(\tau,t) \begin{bmatrix} x \\ y \end{bmatrix}$$

where $P(\cdot)$ is a 2n x 2n symmetric matrix. Then

$$V_x^\circ(x,\tau,y,t) = 2(P_{11}(\tau,t)x + P_{12}(\tau,t)y)$$

$$V_y^\circ(x,\tau,y,t) = 2(P_{12}(\tau,t)x + P_{22}(\tau,t)y),$$

and we are naturally led to the condition:

$$0 = \begin{bmatrix} x \\ y \end{bmatrix}' \{\frac{\partial P}{\partial \tau}(\tau,t) + \frac{\partial P}{\partial t}(\tau,t) + P(\tau,t)A(\tau,t) + A'(\tau,t)P(\tau,t)$$

$$+ P(\tau,t)B(\tau,t)P(\tau,t)\}\begin{bmatrix} x \\ y \end{bmatrix},$$

with

$$A(\tau,t) = \left[\begin{array}{c|c} A_p(\tau) & 0 \\ \hline 0 & A_e(t) \end{array}\right] \qquad B(\tau,t) = \left[\begin{array}{c|c} -B_p(\tau)R_p^{-1}(\tau)B_p'(\tau) & 0 \\ \hline 0 & B_e(t)R_e^{-1}(t)B_e'(t) \end{array}\right]$$

$$\text{and } P(\tau,t) = \left[\begin{array}{c|c} P_{11}(\tau,t) & P_{12}(\tau,t) \\ \hline P_{12}(\tau,t) & P_{22}(\tau,t) \end{array}\right].$$

Since the condition holds for arbitrary x and y, the following generalized matrix Riccati equation must be satisfied over the τ-t plane with $\tau \leq t \leq T$:

$$\frac{\partial P(\tau,t)}{\partial \tau} + \frac{\partial P(\tau,t)}{\partial t} + P(\tau,t)A(\tau,t) + A'(\tau,t)P(\tau,t)$$

$$+ P(\tau,t)B(\tau,t)P(\tau,t) = 0 \qquad (8)$$

subject to:

$$P(\tau,t)\Bigg|_{\substack{\tau=T-\sigma\\ t=T}} = \left[\begin{array}{c} \Phi'_p(T,T-\sigma) \\ \hline -I \end{array}\right] [F^{-1} + SR_p^{-1}S^*]^{-1} [\Phi_p(T,T-\sigma) \mid -I]. \quad (9)$$

From the sufficiency theorem for DGWITL [1] we conclude that the non-realizable optimal exploitation strategy for the pursuer is

$$u^\circ(x,\tau,y,t) = -R_p^{-1}(\tau)B_p'(\tau)[P_{11}(\tau,t)x + P_{12}(\tau,t)y],$$

the DGWITL optimal strategy for the evader is:

$$v^\circ(x,\tau,y,t) = R_e^{-1}(t)B_e'(t)[P_{12}(\tau,t)x + P_{22}(\tau,t)y],$$

and the potential value is

$$V^\circ(x,\tau,y,t) = \begin{bmatrix} x \\ y \end{bmatrix}' P(\tau,t) \begin{bmatrix} x \\ y \end{bmatrix}$$

in the neighborhood of (T-σ,T) in which the solution to (8) is positive definite. Although the quadratic payoff-linear control structure of the solution has been preserved, it is important to note that the generalized Riccati equation is defined on two temporal parameters, while the potential value is defined on the spatial variables and their associated temporal parameters. This point is emphasized by Figure 1, which illustrates the so-called "data reference plane" for this problem. The shaded region of the plane below $\sigma=0(\tau=t)$ is that region over which (4) must be solved. The line $t=T$ is the locus on which the potential terminal payoff provides boundary data for $V^\circ(\cdot)$ over the shaded region. If the DG with zero time lag has already been solved, it may be used to provide boundary data along $\sigma=0$. The behavior of V° along a line of fixed t with $\tau \leq t$ is the "aging data problem," a problem of practical interest.

In general, the generalized HJE over the τ-t plane must be converted to a "delayed argument" partial differential equation in t along the line $\tau = t-\sigma$. So, we must convert the generalized Riccati equation in τ and t into a Riccati equation in t with parameter σ appearing in delayed arguments of the time varying matrices. That is, we wish to generate solutions in parametric form by solving the GHJE along a fixed-σ locus for a DGWITL in Figure 1. For a fixed σ, let

$$\hat{P}(t;\sigma) \triangleq P(\tau,t)\Big|_{\tau=t-\sigma}$$

It follows directly that the solution to (8) along the line $\tau = t-\sigma$ with boundary condition (9) must satisfy:

$$\frac{d}{dt}\hat{P}(t;\sigma) + \hat{P}(t;\sigma)A(t-\sigma,t) + A'(t-\sigma,t)\hat{P}(t;\sigma) + \hat{P}(t;\sigma)B(t-\sigma,t)\hat{P}(t;\sigma) = 0 \qquad (10)$$

subject to:

$$\hat{P}(T;\sigma) = \left[\begin{array}{c} \Phi_p'(T,T-\sigma) \\ \hline -I \end{array}\right][F^{-1} + SR_p^{-1}S^*]^{-1}[\Phi_p(T,T-\sigma) \mid -I] \qquad (12)$$

To indicate the solution along $\tau = t-\sigma$ we write:

$$V^o(x,y,t;\sigma) = \begin{bmatrix} x \\ y \end{bmatrix}' \hat{P}(t;\sigma) \begin{bmatrix} x \\ y \end{bmatrix}$$

$$u^o(x,y,t;\sigma) = -R_p^{-1}(t-\sigma)B_p'(t-\sigma)[\hat{P}_{11}(t;\sigma)x + \hat{P}_{12}(t;\sigma)y]$$

$$v^o(x,y,t;\sigma) = R_e^{-1}(t)B_e'(t)[\hat{P}_{12}(t;\sigma)x + \hat{P}_{22}(t;\sigma)y]$$

Figure 2 illustrates the block diagram for the evader's DGWITL feedback control law. For additional emphasis, we write $x_p(\cdot) = x(\cdot)$ and $x_e(\cdot) = y(\cdot)$. While the linear feedback structure has been preserved, the DGWITL control law is distinct from the control law for the game with zero time lag. The important difference is that the control law uses the pursuer's delayed state vector and Riccati gain function obtained from the solution to a Riccati equation with delayed arguments.

As an alternate scheme, one can define

$$\tilde{P}(t;\sigma) = \left[\begin{array}{c|c} \Phi_p'(t-\sigma,t) & 0 \\ \hline 0 & -I \end{array}\right] \hat{P}(t;\sigma) \left[\begin{array}{c|c} \Phi_p(t-\sigma,t) & 0 \\ \hline 0 & -I \end{array}\right]$$

Then:

$$\begin{bmatrix} u^o(\cdot) \\ v^o(\cdot) \end{bmatrix} = \left[\begin{array}{c|c} -R_p^{-1}(t-\sigma)B_p'(t-\sigma)\Phi_p'(t,t-\sigma) & 0 \\ \hline 0 & R_e^{-1}(t)B_e'(t) \end{array}\right] \tilde{P}(t;\sigma) \left[\begin{array}{c} \Phi_p(t,t-\sigma)x(t-\sigma) \\ \hline y(t) \end{array}\right]$$

where $\tilde{P}(\cdot;\sigma)$ satisfies the Riccati equation in the control law block diagram of Figure 3. This representation of $v^o(\cdot)$ exposes the "simple predictor" portion of the control law, i.e. $\tilde{x}_p(t) = \Phi_p(t,t-\sigma)x_p(t-\sigma)$. The evader can form $\tilde{x}_p(t)$ with certainty, since it is determined solely by P's dynamics. This control law also requires the solution of a Riccati equation with delay arguments.

5. SUMMARY

A presentation of a portion of the theory of differential games with information time lag indicated that the potential value satisfied a generalized Hamilton-Jacobi equation over a portion of the τ-t data reference plane. This equation contains temporal parameters τ and t, the times associated with the spatial data observed by the evader. For solution purposes, we showed that the GHJE can easily be converted to a p.d.e. with temporal parameter t along the line $\tau = t - \sigma$. Lastly, the well-known regulator game was solved to illustrate the application of the theory of DGWITL. The resulting feedback controller required solution of a matrix Riccati equation with delay arguments.

Reference

1. Ciletti, M. D., "New Results in the Theory of Differential Games with Information Time Lag", Journal of Optimization Theory and Applications, Vol. 8, No. 4, 1971

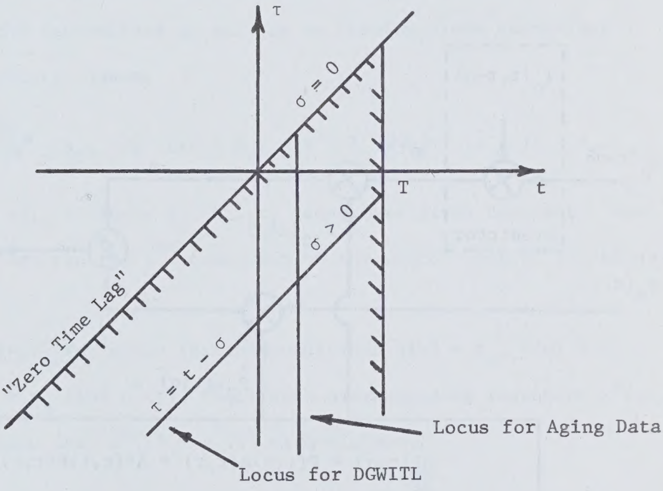


Figure 1. Data Reference Plane

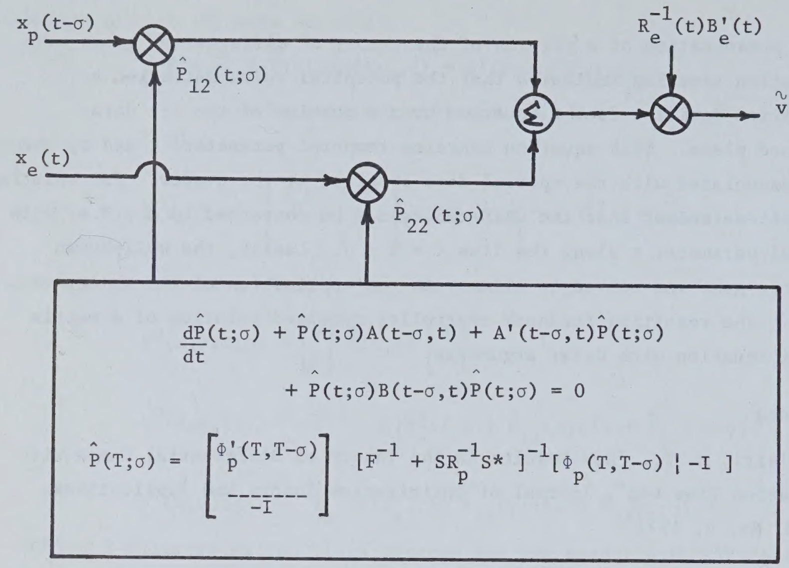

$x_p(t-\sigma)$
$R_e^{-1}(t)B_e'(t)$
$\hat{P}_{12}(t;\sigma)$
Σ
$\tilde{v}$
$x_e(t)$
$\hat{P}_{22}(t;\sigma)$
$\frac{d\hat{P}}{dt}(t;\sigma) + \hat{P}(t;\sigma)A(t-\sigma,t) + A'(t-\sigma,t)\hat{P}(t;\sigma) + \hat{P}(t;\sigma)B(t-\sigma,t)\hat{P}(t;\sigma) = 0$
$\hat{P}(T;\sigma) = \left[\begin{array}{c}\Phi_p'(T,T-\sigma)\\ \hline -I\end{array}\right] [F^{-1} + SR_p^{-1}S^*]^{-1}[\Phi_p(T,T-\sigma) \mid -I]$

Figure 2. Evader DGWITL Control Law

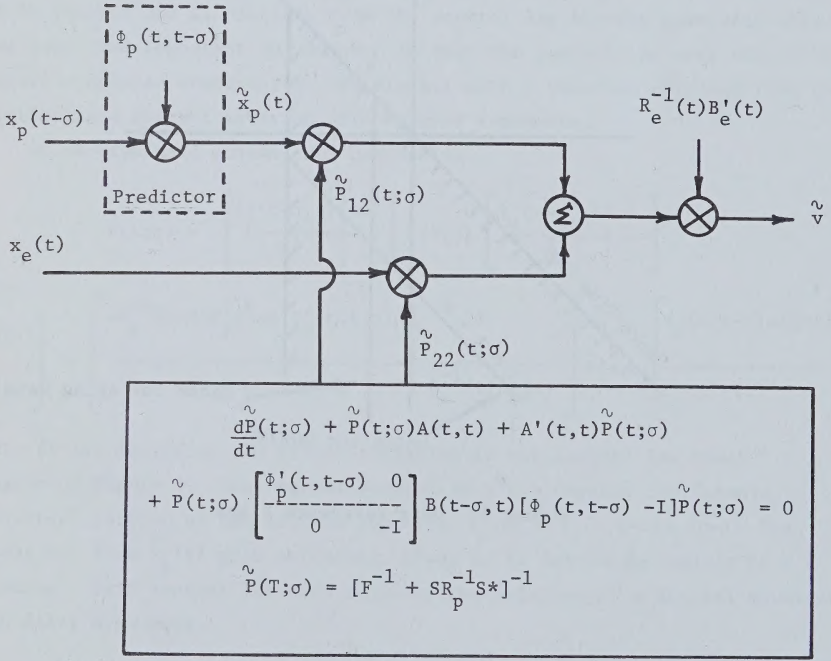

$\Phi_p(t,t-\sigma)$
$\tilde{x}_p(t)$
$x_p(t-\sigma)$
Predictor
$R_e^{-1}(t)B_e'(t)$
$\tilde{P}_{12}(t;\sigma)$
Σ
$\tilde{v}$
$x_e(t)$
$\tilde{P}_{22}(t;\sigma)$
$\frac{d\tilde{P}}{dt}(t;\sigma) + \tilde{P}(t;\sigma)A(t,t) + A'(t,t)\tilde{P}(t;\sigma) + \tilde{P}(t;\sigma)\left[\begin{array}{cc}\Phi_p'(t,t-\sigma) & 0\\ \hline 0 & -I\end{array}\right]B(t-\sigma,t)[\Phi_p(t,t-\sigma)\ -I]\tilde{P}(t;\sigma) = 0$
$\tilde{P}(T;\sigma) = [F^{-1} + SR_p^{-1}S^*]^{-1}$

Figure 3. Predictor Decomposition of Evader Control Law

ON RENDEZVOUS AND EVASION OF LINEAR SYSTEMS WITH MULTIPLE CONSTRAINTS ON CONTROLLERS

DONG HAK CHYUNG
Department of Electrical Engineering
University of Iowa
Iowa City, Iowa 52240

ABSTRACT

Problems of rendezvous and evasion of linear systems with multiple constraints on the controllers are investigated.

For simplicity consider two normal controllable linear time invariant vector systems (see [4] for terminology)

$$\dot{x}(t) = A\,x(t) + B\,u(t)$$

$$\dot{y}(t) = C\,y(t) + D\,v(t)$$

The control functions u(t) and v(t) are absolutely continuous functions with bounds on the derivatives as well as on the functions themselves. Again, for simplicity, assume

$$|u^i(t)| \le a_i, \quad |\dot{u}^i(t)| \le b_i, \quad |v^i(t)| \le c_i, \quad |\dot{v}^i(t)| \le d_i$$

for all t for all i, where a_i, b_i, c_i and d_i are given constants, and the superscript denotes the i^{th} component of the vector control functions.

Rendezvous Problem: For given initial conditions $x(o) = x_o$, $y(o) = y_o$, $u(o) = u_o$, $v(o) = v_o$ find $u^*(t)$, $v^*(t)$ with corresponding responses $x^*(t)$, $y^*(t)$ on [0,T] such that $x^*(T) = y^*(T)$ with minimum T.

Evasion Problem: For given initial conditions $x(o) = x_o$, $y(o) = y_o$, $u(o) = u_o$, $v(o) = v_o$ and finite time T find $u^*(t)$, $v^*(t)$ with corresponding responses $x^*(t)$, $y^*(t)$ on [0, T] such that the distance $|x^*(T) - y^*(T)|$ is maximized.

By examining the sets of attainability in a similar manner as in [3] the following results can be obtained.

Rendezvous Problem: Let $u^*(t)$ and $v^*(t)$ on [0, T] be a solution pair of the problem. Then there exist nontrivial solutions p(t) and q(t) of

$$\dot{p}(t) = -p(t)\,A, \quad \dot{q}(t) = -q(t)\,C$$

with the endpoint condition p(T) = -q(T) such that

$$\int_o^T p(t)\,B\,u^*(t)\,dt = \max_u \int_o^T p(t)\,B\,u(t)\,dt$$

$$\int_o^T q(t)\,D\,v^*(t)\,dt = \max_v \int_o^T q(t)\,D\,v(t)\,dt$$

for all allowable control functions u(t) and v(t) on [0, T]. Furthermore the above maximal conditions imply that either $|u^{i*}(t)| = a_i$ or $|\dot{u}^{i*}(t)| = b_i$ and either $|v^{i*}(t)| = c_i$ or $|\dot{v}^{i*}(t)| = d_i$ for all i and all t on [0, T].

Evasion Problem: Let $u^*(t)$ and $v^*(t)$ be a solution pair of the problem. Then the results obtained in the above rendezvous problem are also satisfied by the control pair $u^*(t)$ and $v^*(t)$ except that the endpoint condition p(T) = -q(T) should be replaced by the condition $p(T) = -q(T) = x^*(T) - y^*(T)$.

REFERENCES

1. S.S.L. Chang, "Minimum Time Control with Multiple Saturation Limits," IEEE Trans. on Automatic Control, Vol. AC-8, No.1, January 1963.
2. D.H. Chyung, "On a Class of Pursuit-Evasion Game," IEEE Trans. on Automatic Control, Vol. AC-15, No. 4, August 1970.

3. D.H. Chyung, "On a Pursuit-Evasion Problem with Multiple Constraints on the Controllers," IEEE Trans. on Automatic Control, Vol. AC-16, No. 5 October 1971.

4. E.B. Lee and L. Markus, "Foundations of Optimal Control Theory," Wiley, New York, 1967.

OPTIMUM CONTROL IN A DISTRIBUTED SYSTEM UNDER CONFLICT

MARSHALL C.Y. KUO
Newark College of Engineering
Newark, New Jersey

ABSTRACT

This paper deals with terminal control and time control in a linear distributed parameter system for two players who have conflicting goals. Analytical solution is obtained by the method of functional analysis.

INTRODUCTION

Differential game problems with quadratic criterion in distributed parameter systems have appeared in the literature [1, 2]. The purpose of this paper is to study the optimal terminal control and time control of a class of distributed parameter systems under conflict.

Let R be the region interior to a solid bounded by a closed interval ∂R so that

$$R = \{ z : 0 < z < 2\pi \}$$

and

$$\overline{R} = R \cup \partial R = \{ z : 0 \le z \le 2\pi \}. \tag{1.1}$$

Let $\tau = (0,T]$ and $Q = R \times \tau$. Then, the temperature $x(z,t)$ in a slab with initial temperature $v(z)$, when the face $z = 0$ is held at zero and the face $z = 2\pi$ is insulated, may be described by the following set of equations:

$$\frac{\partial x}{\partial t} = k \frac{\partial^2 x}{\partial z^2} + u(z,t), \qquad (z,t) \in Q \tag{1.2}$$

$$\left.\begin{aligned} x(+0,t) &= 0 \\ \frac{\partial x}{\partial z}(2\pi-0,t) &= 0 \end{aligned}\right\} \qquad t \in \tau \tag{1.3}$$

$$x(z, +0) = v(z), \qquad z \in R \tag{1.4}$$

where $u(z,t)$ is a source of heat in Q satisfying the constraint

$$\int_0^T \int_0^{2\pi} |u(z,t)|^2 \, dz \, dt < \infty \quad ; \tag{1.5}$$

k is the thermal conductivity; and $v(z)$ is the initial temperature independent of $u(z,t)$.

Using the method of separation of variables we can show that the solution of homogeneous form of (1.2-1.4) is given by

$$x(z,t) = \frac{1}{\pi} \sum_{i=1}^{\infty} e^{-m_i^2 kt} \sin m_i z \int_0^{2\pi} v(z) \sin m_i z \, dz \qquad (1.6)$$

where $m_i = \frac{2i-1}{4}$, $i = 1,2,\ldots.$. With forcing function $u(z,t)$ the solution of nonhomogeneous equation of (1.2-1.4) has the form [2]

$$x(z,t) = \underline{Z}^T(z)\Phi(t,0)\underline{c} + \underline{Z}^T(z) \int_0^t \Phi(t,s)\underline{g}(s)ds \qquad (1.7)$$

where

$$\underline{Z}(z) = \text{col.}(Z_1(z), Z_2(z), \ldots), \quad Z_i(z) = \sqrt{\frac{1}{\pi}} \sin m_i z \qquad (1.8)$$

$$\underline{c} = \text{col.}(c_1, c_2, \ldots), \quad c_i = \sqrt{\frac{1}{\pi}} \int_0^{2\pi} v(z) \sin m_i z \, dz \qquad (1.9)$$

$$\underline{g} = \text{col.}(g_1, g_2, \ldots), \quad g_i = \sqrt{\frac{1}{\pi}} \int_0^{2\pi} u(z,t) \sin m_i z \, dz \qquad (1.10)$$

$i = 1,2,\ldots$

$$\Phi(t,s) = \begin{bmatrix} e^{-km_1^2(t-s)} & & 0 \\ & e^{-km_2^2(t-s)} & \\ 0 & & \ddots \end{bmatrix} \qquad (1.11)$$

and $\underline{Z}^T$ denotes the transpose of the vector $\underline{Z}$.

Let $L_2(0,2\pi)$ be the real Hilbert space of square integrable functions defined on the closed interval $[0,2\pi]$ with the inner product defined by

$$(x,y) = \int_0^{2\pi} x(z)y(z)dz \qquad (1.12)$$

for $x,y \in L_2(0,2\pi)$.

Let W be the real Hilbert space of all functions, measurable on Q, with the inner product defined by

$$\langle x,y \rangle = \int_0^t \int_0^{2\pi} x(z,s)y(z,s)dzds \qquad (1.13)$$

for each fixed $t \in (0,T]$ and $x,y \in W$.

Define

$$p_i(z,t,s) = \sqrt{\frac{1}{\pi}} e^{-km_i^2(t-s)} \sin m_i z, \quad i=1,2,\ldots, \qquad (1.14)$$

and consider p_i as an element in $L_2(o,2\pi)$ by fixing both t and s, and p_i as an element in W by fixing t only. Then, it follows from (1.12-1.14) that (1.7) may be written as

$$x(z,t) = \sum_{i=1}^{\infty} Z_i(z)\{(p_i,v)(t) + \langle p_i, u\rangle(t)\} = \sum_{i=1}^{\infty} Z_i(z)T_i(t) \quad (1.15)$$

where $T_i(t) = (p_i,v)(t) + \langle p_i,u\rangle(t)$.

For each fixed t and z equation (1.15) may be identified as a mapping from ℓ_2 into E (one dimensional Euclidean space). Let us define

$$G_1(v)(z,t) = \sum_{i=1}^{\infty} Z_i(z)(p_i,v)(t) \quad (1.16)$$

$$G_2(u)(z,t) = \sum_{i=1}^{\infty} Z_i(z)\langle p_i, u\rangle(t). \quad (1.17)$$

Then, from (1.15-1.17), the solution of (1.2-1.4) in operator form may be written as

$$x(z,t) = G_1(v)(z,t) + G_2(u)(z,t) \overset{\Delta}{=} G(u,v)(z,t). \quad (1.18)$$

In view of (1.15-1.17) we have the following Theorem 1.
Theorem 1: G(u,v), defined by (1.18), is a bounded linear operator of $W \times L_2$ into E, where $(u,v) \in W \times L_2$ and $x \in E$.
Proof: By inspection G(u,v) defined by (1.16-1.18) is a linear operator. Let the norms of x in E and (u,v) in $W \times L_2$ be, respectively, denoted by $|x|$ and $\|(u,v)\|$. Furthermore, define an inner product on $W \times L_2$ by

$$[(u_1, v_1), (u_2, v_2)]_{W \times L_2} = \langle u_1, u_2\rangle_W + (v_1, v_2)_{L_2}.$$

Then [3], we have

$$\|(u,v)\|^2 = \|u\|_W^2 + \|v\|_{L_2}^2.$$

It follows from (1.15-1.18) that

$$|x(z,t)| = |G(u,v)| = \left|\sum_{i=1}^{\infty} Z_i(z)\{(p_i,v)(t) + \langle p_i,u\rangle(t)\}\right|$$

$$\leq \sum_{i=1} |Z_i(z)|\{|(p_i,v)(t)| + |\langle p_i,u\rangle(t)|\}. \quad (1.19)$$

Using Schwarz inequality, for each fixed t and i, yields

$$|(p_i,v)(t)| = \left|\int_0^{2\pi} p_i(z,t,0)v(z)dz\right|$$

$$\leq \{|\int_0^{2\pi} |p_i(z,t,0)|^2 dz\}^{1/2} \{\int_0^{2\pi} |v(z)|^2 dz\}^{1/2}$$

$$= \|p_i(t)\|_{L_2} \|v\|_{L_2}.$$

Similarly, the following inequality, with the aid of (1.13-1.14),

$$| \langle p_i , v \rangle \, (t) | \le \| p_i(t) \|_W \, \| u \|_W$$

is held. Define

$$Y_i = \sup_{t \varepsilon [0,T]} \{ \| p_i(t) \|_{L_2} , \| p_i(t) \|_W \}$$

and

$$I_i = \sup_{z \varepsilon [0,2\pi]} \{ Z_i(z) , Y_i \}.$$

Then, (1.19) reduces to

$$|G(u,v)| \le \sum_{i=1}^{\infty} |Z_i(z)| \, |Y_i| \, \{ \| u \|_W + \| v \|_{L_2} \}$$

$$\le \sum_{i=1}^{\infty} |I_i|^2 \{ \| u \|_W + \| v \|_{L_2} \}$$

$$\le M \, \| (u,v) \|^{*} \text{ for each fixed z and t,}$$

where $\frac{M}{2} = \sum_{i=1}^{\infty} |I_i|^2 = \| I \|^2_{\ell_2}$ (norm in ℓ_2 space) and $I = (I_1, I_2, \ldots)$.

This completes the proof that G is a bounded linear operator from $W \times L_2 \to E$.

2. STATEMENT OF PROBLEMS

Let us consider $u \varepsilon W$ and $v \varepsilon L_2$ as two controls operated by two players (Man and Nature) who have conflicting goals. Player I (Man) strives to minimize a specified pay off $J(u,v)$ while player II (Nature), on the contrary, strives to maximize $J(u,v)$. Let the set S be defined by

$$S = \{ (u,v) : |u| \le L, \ |v| \le M, \ \text{and } x = G(u,v) \} \qquad (2.1)$$

where L and M are constants greater than zero.

---

*For two numbers a and b, with $p \ge 1$, $|a + b|^p \le 2^p (|a|^p + |b|^p)$. In particular, $p = 2$, then, $|a+b| \le 2 \sqrt{|a|^2 + |b|^2}$.

If $x(z,t)$ is an arbitrary element of E, with each fixed z and t, from the set S find the saddle point [4,5] of the following game problems:

Problem 1:

$$J_1(u^o,v^o) = \min_{|u| \leq L} \max_{|v| \leq M} \{J_1(u,v)\}$$

$$= \min_{|u| \leq L} \max_{|v| \leq M} \{x(C,T) - d\} \tag{2.2}$$

where $x(C,T)$ is free but d and T are fixed.

Problem 2:

$$J_2(u^o,v^o) = \min_{|u| \leq L} \max_{|v| \leq M} \{T(u\ v)\} = T(u^o,v^o) \tag{2.3}$$

where $T = T(u,v)$ is free and

$$x(C,T) = d \tag{2.4}$$

is fixed.

3. SOLUTIONS TO PROBLEM 1 AND PROBLEM 2

The solutions to Problem 1 and Problem 2 are summarized in the following Theorem 2 and 3, respectively.

Theorem 2: Let r be the element of E*, the dual space of E, and Ψ and ψ be, respectively, defined by

$$\Psi = G_1^* r \quad \text{and} \quad \psi = G_2^* r \tag{3.1}$$

where G_1^* (G_2^*) is the adjoint of G_1 (G_2). Then, the pair (u^o,v^o) for Problem 1 satisfies

$$rG_1(v^o) = \|\Psi\|_{L_2} M \tag{3.2}$$

$$rG_2(u^o) = - \|\psi\|_W L \tag{3.3}$$

and

$$v^o(z) = \frac{M}{\|\Psi\|_{L_2}} |\Psi| \operatorname{sgn}\{\Psi\} \tag{3.4}$$

$$u^o(z,t) = - \frac{L}{\|\psi\|_W} |\psi| \operatorname{sgn}\{\psi\}. \tag{3.5}$$

The value of game for Problem 1 is given by

$$J_1(u^o,v^o) = f(r^o) \tag{3.6}$$

where $f(r^o) = \| \Psi^o \|_{L_2} M - \| \psi^o \|_W L - r^o d$,

$$\Psi = G_1^* r^o,$$
$$\psi = G_2^* r^o,$$

and r^o is the scalar obtained by maximizing the functional $f(r)$ subject to the constraint $|r| \leq 1$. The proof of Theorem 2, which is similar to that of Theorem 5 as given in Reference [5], is omitted.

Let us introduce a function

$$h(v,T,t_o) = \min_{|u| \leq L} \max_{|r| \leq 1} \{r[x(C,T) - d\ ]\}$$

$$= \max_{|r| \leq 1} \{rG_1(v) - \| \psi \|_W L - rd]\}. \qquad (3.7)$$

The solution to problem 2 is summarized in the following Theorem 3.

Theorem 3: If the following three conditions are satisfied

(1) Ψ and ψ, defined by (3.1), are continuously increasing functions of T,

(2) $h(v,T,t_o) \leq 0$ for $v \varepsilon S$, (3.8)

(3) there exists a constant k, $0 < k < \infty$, such that

$$k \geq \| \psi \|_W L + rd \qquad (3.9)$$

for all $r \varepsilon E^*$ and $T(\infty > T > t_o)$, then the value of game $T = T(u^o,v^o)$ for Problem 2 can be computed from

$$\| \Psi^o \|_{L_2} M - \| \psi^o \|_W L - r^o d = 0 \qquad (3.10)$$

where r^o is the solution of (3.7) with $T = T(u^o,v^o)$ and $v = v^o$. The optimal pair (u^o,v^o) satisfies (3.2) and (3.3), respectively.

Proof: From condition (2) with $r = r^o$ we have

$$h(v,T,t_o) = r^o G_1(v) - \| \psi^o \|_W L - r^o d \leq 0 \qquad (3.11)$$

where

$$r^o G_2(u^o) = \min_{|u| \leq L} \{r^o G_2(u)\} = - \| G_2^* r^o \|_W L = - \| \psi^o \|_W L. \qquad (3.12)$$

The least $T = T(v) = \min_{|u| \leq L} T(u,v)$ satisfying (3.11) is equal to the minimal time that x can reach to a point d with the element v.

It follows from (3.9) and (3.11) that

$$k \geq \|\psi^o\|_W L + r^o d \geq r^o G_1(v) \text{ for all } v \varepsilon S. \tag{3.13}$$

In particular, (3.13) is held for $v = v^o$, where v^o is the vector satisfying the following equation

$$r^o G_1(v) = \max_{|v| \leq M} \{r^o G_1(v)\} = \|G_1^* r^o\|_{L_2} M = \|\Psi^o\|_{L_2} M. \tag{3.14}$$

Combining the last two equations gives

$$k \geq \|\psi^o\|_W L + r^o d \geq \|\Psi^o\|_{L_2} M. \tag{3.15}$$

From (3.15) the largest $T = \max_{|v| \leq M} T(v) = T(u^o,v^o)$ can be

achieved only when

$$\frac{k}{M} = \|\Psi^o\|_{L_2} \tag{3.16}$$

in view of condition (1). However, (3.15) and (3.16) imply that $T = T(u^o,v^o)$ must satisfy

$$k = \|\psi^o\|_W L + r^o d = \|\Psi^o\|_{L_2} M \tag{3.17}$$

where r^o is the solution of (3.7) with $T = T(u^o,v^o)$ and $v = v^o$. The optimal u^o and v^o can be calculated from (3.12) and (3.14) or (3.3) and (3.2), respectively. With the aid of (3.7) it is easily shown that (3.17) satisfies the terminal condition (2.4) at $T = T(u^o,v^o)$.

4. CONCLUSION

Game problems in a class of distributed parameter systems for the terminal control and time control on function space have been studied. It is interesting to observe that the saddle point for Problem 1 coincides with that for Problem 2. The theory developed in this paper is not only applicable to a linear diffusion process but also to other types of linear distributed parameter systems.

REFERENCES

1. Hemming, F., and V.D. Vandelinde, "On Some Aspects of Pursuit Evasion Games," Proceedings of the Fourth Annual Princeton Conference on Information Sciences and Systems, 1970, pp. 445
2. Porter, W.A., "On Function Space Pursuit-Evasion Games," SIAM J. on Control. Vol. 5, No. 4, 1967, pp. 555-574.
3. Kuo, M.C.Y., "Minimum Energy Problems in Hilbert Function Space," Journal of the Franklin Institute, Vol. 283, No. 1, 1967, pp. 38-54.
4. Berkovitz, L.D., "A Survey of Differential Games," Mathematical Theory of Control, edited by A.V. Balakrishnan and L.W. Neustadt, Academic Press, 1967, pp. 342-372.
5. Kuo, M.C.Y., "Differential Games on Function Space," Journal of the Franklin Institute, Vol. 289, No. 6, 1970, pp. 411-426.

A TWO PERSON ZERO SUM STOCHASTIC GAME THAT ADMITS A MAXIMIN POLICY DEFINED BY A GAUSSIAN MEASURE

TAMER BASAR and MAX MINTZ
Department of Engineering and Applied Science, Yale University
New Haven, Connecticut, 06520

ABSTRACT

Optimal strategies are derived for the following two person zero sum stochastic game:

Let $z = x + v + w$, where $x \in N[\underline{o},Q]$, $w \in N[\underline{o},R]$, and $v \in E^n$. Player 1 observes the random vector z and seeks a minimax policy u(z) under the quadratic payoff function defined by

$$L(u,v) = [u + x + v]'C[u + x + v] + u'D_1u - v'D_2v.$$

The vector v, which is unknown to player 1, is under the control of player 2. Player 2 does not make any observations, but seeks a maximin policy for v, by picking v to be a random vector statistically independent of the random vectors x and w.

An admissible minimax policy for player 1 is shown to be a linear transformation on z, and a maximin policy for player 2 (a probability distribution for v which is least favorable to player 1) is shown to be a normal distribution with zero mean and covariance Λ^*, where Λ^* is determined by the solution of a certain nonlinear matrix equation.

The stochastic game discussed in this paper can be viewed as a special case of a discrete stochastic pursuit-evasion game of the linear quadratic gaussian (LQG) variety where one player (the pursuer) makes noisy state measurements and the other player (the evader) makes no state measurements. In this regard, the major contribution of this paper is a rigorous proof of the existence of a maximin policy, defined by a gaussian measure, for the evader in this specific game; this proof can be extended to include the general LQG pursuit-evasion game.

I. PROBLEM STATEMENT

A two person zero sum stochastic game is posed as follows:

Let z denote a normal random vector defined by

$$z = x + v + w \quad , \tag{1}$$

where $x \in N[\underline{o},Q]$, $w \in N[\underline{o},R]$, $v \in E^n$.

It is assumed that x and w are statistically independent, and that $Q \geq \underline{0}$, and $R > \underline{0}$. Player 1 observes the random vector z and seeks a minimax policy u(z) under the quadratic payoff function L(u,v) defined by

$$L(u,v) = [u + x + v]'C[u + x + v] + u'D_1u - v'D_2v \quad , \tag{2}$$

where $C \geq \underline{0}$, $D_1 > \underline{0}$, and $D_2 > \underline{0}$. The vector v, which is under the control of player 2, is unknown to player 1.

Player 2 does not make any observations, but seeks a maximin policy for v, by picking v to be a random vector which is statistically independent of the random vectors denoted by x and w. This maximin policy can also be identified as a probability distribution which is least favorable to player 1. Both players know the statistics of the indicated normal random vectors x and w.

Associated with the payoff function defined by (2) is the conditional payoff function $\tilde{L}(u,v)$ defined by

$$\tilde{L}(u(z),v) = E[L(u(z),v)|v] \quad , \tag{3}$$

where E[.] denotes the expectation operation taken over the prior statistics of x and w; and $u(.) \in \Sigma$ where Σ is the space of all Borel measureable functions mapping E^n into E^n.

Definition of a minimax policy for player 1:

In the context of the previous notation, a policy $u(z) \in \Sigma$ is said to be minimax, if for all policies $u_o(z) \in \Sigma$, the following inequality holds:

$$\sup_{v \in E^n} \tilde{L}(u(z),v) \leq \sup_{v \in E^n} \tilde{L}(u_o(z),v) \quad . \tag{4a}$$

Definition of a least favorable distribution for v:

A distribution λ is said to be least favorable if for all distributions μ,

$$\inf_{u(.)\in\Sigma} \int_{E^n} \tilde{L}(u,v)d\lambda \geq \inf_{u(.)\in\Sigma} \int_{E^n} \tilde{L}(u,v)d\mu \quad . \tag{4b}$$

II. MAIN RESULTS

Define $\Gamma(\Lambda)$ by

$$\Gamma(\Lambda) = [I + K(\Lambda)]'C[I + K(\Lambda)] + K'(\Lambda)D_1K(\Lambda) - D_2 \quad , \tag{5a}$$

where

$$K(\Lambda) \triangleq -[C + D_1]^{-1}C[\Lambda + Q][\Lambda + Q + R]^{-1} \quad , \tag{5b}$$

and

$$\Lambda \geq \underline{0} \quad . \tag{5c}$$

Let Λ^* denote any nonnegative definite solution to the following nonlinear matrix equation:

$$\Gamma(\Lambda^*)\Lambda^* = \underline{0} \quad , \tag{6a}$$

such that

$$\Gamma(\Lambda^*) \leq \underline{0} \quad . \tag{6b}$$

Further assume that the following condition is satisfied,

$$C[C + D_1]^{-1}D_1 - D_2 < \underline{0} \quad . \tag{7}$$

THEOREM

Assume that the condition given by (7) holds.

i. If $\Gamma(\Lambda)$ is defined by (5a), (5b), and (5c), then there exists at least one symmetric matrix $\Lambda^* \geq \underline{0}$, which satisfies (6a) and (6b).

ii. Under the quadratic payoff function defined by (2), an admissible minimax policy for player 1 is given by

$$u^*(z) = -[C + D_1]^{-1}C[\Lambda^* + Q][\Lambda^* + Q + R]^{-1}z \quad , \tag{8}$$

where Λ^* is defined by (6a) and (6b).

iii. Under the quadratic payoff defined by (2), an optimum mixed strategy (maximin policy) for player 2 is to pick

$$v^* \in N[\underline{o},\Lambda^*] \quad , \tag{9}$$

where Λ^* is defined by (6a) and (6b).

A proof of part (i) of the Theorem is given in Appendix I. A proof of parts (ii) and (iii) of the Theorem is developed by using Lemma 1, which is given in section III.

III. A LEAST FAVORABLE PRIOR DISTRIBUTION FOR v

The following lemma provides us with a useful relationship between an optimum choice for u (a minimax policy for player 1) and a least favorable prior distribution for v (a maximin policy for player 2).

Let λ denote a prior probability measure for v. Further, let $u_\lambda(z)$ denote a bayes policy for player 1 with respect to λ, and the payoff function defined by (2).

Lemma 1

Assume that $\max_{v \in E^n} \tilde{L}(u_\lambda(z),v)$ exists, and denote its value by m_λ. Let $\tau_\lambda = \{v: \tilde{L}(u_\lambda(z),v) = m_\lambda\}$.

If $\lambda[\tau_\lambda] = 1$, then the bayes policy $u_\lambda(z)$ is a minimax policy for player 1, and λ is a least favorable prior probability measure for v.

A proof of Lemma 1 is given in [4], [8].

Lemma 1 is used as follows:

Let λ_o denote a prior probability measure for v such that $v \in N[\underline{o},\Lambda]$. For this choice of λ_o, a bayes policy $u_{\lambda_o}(z)$ is

$$u_{\lambda_o}(z) = -[C + D_1]^{-1} C\, E[x + v|z] \quad ; \tag{10a}$$

which follows by completing the square in expression (3). When $v \in N[\underline{o},\Lambda]$,

$$E[x + v|z] = [\Lambda + Q][\Lambda + Q + R]^{-1} z \quad ,$$

and expression (10a) becomes

$$u_{\lambda_o}(z) = K(\Lambda) z \quad , \tag{10b}$$

where $K(\Lambda)$ is defined as before by

$$K(\Lambda) = -[C + D_1]^{-1} C[\Lambda + Q][\Lambda + Q + R]^{-1} \quad . \tag{5b}$$

The corresponding conditional payoff is

$$\tilde{L}(u_{\lambda_o}(z),v) = v'\Gamma(\Lambda)v + \mathrm{Tr}[K'(\Lambda)(C + D_1)K(\Lambda)R] + \mathrm{Tr}[(\Gamma(\Lambda) + D_2)Q] \quad , \tag{11}$$

where $K(\Lambda)$ is defined by (5b), and $\Gamma(\Lambda)$ is defined as before by

$$\Gamma(\Lambda) = [I + K(\Lambda)]'C[I + K(\Lambda)] + K'(\Lambda)D_1K(\Lambda) - D_2 \quad . \tag{5a}$$

In the proof of part (i) of the Theorem, it is shown that a covariance matrix Λ^* always exists such that relations (6a) and (6b) are satisfied, whenever the condition given by (7) holds [See Appendix I]. In what follows we shall show that the pair

$$\lambda^* = N[\underline{o},\Lambda^*] \tag{12a}$$

$$u_{\lambda^*}(z) = K(\Lambda^*)z \tag{12b}$$

satisfies the sufficiency conditions of Lemma 1, and hence prove parts (ii) and (iii) of the Theorem.

If (6a) is satisfied, then all of the eigenvectors of Λ^* with non-zero eigenvalues are in the null-space of $\Gamma(\Lambda^*)$. Thus the null-space of $\Gamma(\Lambda^*)$, which is a Borel set in E^n, receives probability one from the measure λ^*. Further, if (6b) is satisfied, it follows from (11) that τ_{λ^*} is the null-space of $\Gamma(\Lambda^*)$. Hence $\lambda^*[\tau_{\lambda^*}] = 1$.

IV. ADMISSIBILITY

Definition 1

In the context of the notation given in section I, a policy $u(z)$ is said to be better than $u_o(z)$ if $L(u,v) \leq L(u_o,v)$ for all $v \in E^n$, and $L(u,v) < L(u_o,v)$ for at least one $v \in E^n$.

Definition 2

A policy $u(z)$ is said to be admissible if there exists no other policy better than $u(z)$.

The admissibility of the bayes policy $u^*(z)$ defined by (8) is a consequence of the following lemma which is related to the one found in [[3], p.60].

Lemma 2

If for a given prior distribution λ a bayes policy with respect to λ is unique up to equivalance, this bayes policy is admissible. [A policy u_1 is said to be equivalent to a policy u_2 if $L(u_1,v) = L(u_2,v)$ for all $v \in E^n$.

This lemma is used as follows: Let the prior distribution λ for v be given by

$$\lambda^* = N[\underline{o},\Lambda^*] \quad . \tag{12a}$$

Since the payoff function given by (2) is quadratic in u weighted by a positive definite matrix $(C + D_1)$, the bayes policy with respect to λ^* is unique. Hence, by Lemma 2, the minimax policy $u^*(z)$ is admissible.

V. DISCUSSION OF THE SOLUTION TO THE NONLINEAR MATRIX EQUATION $\Gamma(\Lambda^*)\Lambda^* = \underline{0}$

It is stated in part (i) of the Theorem and proven in Appendix I, that if the condition given by (7) holds, then there exists at least one symmetric matrix $\Lambda^* \geq \underline{0}$, which satisfies (6a) and (6b).

Note that, when $K(\Lambda)$, given by (5b), is substituted into $\Gamma(\Lambda)$, given by (5a), after some algebra one obtains,

$$\Gamma(\Lambda) = [\Lambda + Q + R]^{-1}RC[C + D_1]^{-1}CR[\Lambda + Q + R]^{-1} - P \quad , \tag{13a}$$

where

$$P \triangleq D_2 - C[C + D_1]^{-1}D_1 \quad . \tag{13b}$$

Hence, Λ^* satisfies the equation

$$\Gamma(\Lambda^*)\Lambda^* = \underline{0} \quad , \tag{14a}$$

under the constraint

$$\Gamma(\Lambda^*) \leq \underline{0} \quad , \tag{14b}$$

whenever

$$P > \underline{0} \quad , \tag{14c}$$

where $\Gamma(\Lambda)$ and P are defined by (13a) and (13b) respectively.

The determination of Λ^* is dependent on the signs of the eigenvalues of the matrix M defined by

$$M = [Q + R]^{-1}RC[C + D_1]^{-1}CR[Q + R]^{-1} - P \quad . \tag{15}$$

There are three cases to consider:

Case 1. $M \geq \underline{0}$
Case 2. $M \leq \underline{0}$
Case 3. M has at least one positive and one negative eigenvalue.

Case 1

When $M \geq \underline{0}$, a suitable choice of Λ^* is given by

$$\Lambda^* = [\overline{P^{1/2}}]^{-1}[P^{1/2}RC(C + D_1)^{-1}CRP^{1/2}]^{1/2}[P^{1/2}]^{-1} - [Q + R] \quad . \tag{16}$$

Assertion:

If Λ^* is given by (16), then $\Gamma(\Lambda^*) = \underline{0}$ (thus constraint (14b) is satisfied), and $\Lambda^* \geq \underline{0}$ whenever $M \geq \underline{0}$.

The fact that $\Gamma(\Lambda^*) = \underline{0}$ can be verified by substitution. We show that $\Lambda^* \geq \underline{0}$ as follows. Note that $M \geq \underline{0}$ implies

$$[Q + R]M[Q + R] \geq \underline{0} \quad . \tag{17a}$$

And, further, using the expression for M given by (15), equation (17a) implies

$$\begin{aligned} P^{1/2}RC[C + D_1]^{-1}CRP^{1/2} &\geq P^{1/2}[Q + R]P[Q + R]P^{1/2} \\ &= [P^{1/2}(Q + R)P^{1/2}]^2 \quad . \end{aligned} \tag{17b}$$

But since $A \geq B \geq \underline{0}$ implies that $A^{1/2} \geq B^{1/2}$ [Refer to [2]], it follows that (17b) implies that Λ^* given by (16) is nonnegative definite.

Note that, since $\Gamma(\Lambda^*) = \underline{0}$, whenever Λ^* is given by (16), it follows from (11) that the conditional payoff function is constant for all $v \in E^n$. A policy with constant conditional payoff function is often referred to as an equalizer policy [[3],[7]].

Case 2

When $M \leq \underline{0}$, a suitable choice of Λ^* is given by

$$\Lambda^* = \underline{0} \quad , \tag{18a}$$

since $\Gamma(\underline{0}) = M \leq \underline{0}$. Hence, when $M \leq \underline{0}$, an admissible minimax policy for player 1 is given by

$$u^*(z) = -[C + D_1]^{-1}CQ[Q + R]^{-1}z \quad , \tag{18b}$$

and a maximin policy for player 2 is to pick

$$v^* = \underline{o}, \text{ with probability one.} \tag{18c}$$

Case 3

When M has at least one positive and one negative eigenvalue, a suitable choice of Λ^* is made by picking any matrix $\Lambda^* \geq \underline{0}$, that satisfies (14a) and (14b). It is stated in part (i) of the Theorem and proven in Appendix I that there exists at least one such Λ^* whenever $P > \underline{0}$, where P is defined by (13b).

VI. CONCLUSION

In this paper we have proven the existence of a minimax and a maximin policy for a particular two person zero sum stochastic game. An admissible minimax policy for player 1 is shown to be a linear transformation on his noisy observation [See (8)]; a maximin policy for player 2 is shown to be a normal distribution with zero mean and covariance Λ^* [See (9)], where Λ^* is defined by a solution to the nonlinear matrix equation (6a) under the constraint (6b). A separation result is shown to hold for the first player's minimax strategy, that is, player 1's minimax strategy is (i) to form an estimate (the conditional mean) of the vector $[x + v]$ through his observation z (assuming $v = v^*$), and then (ii) to apply the linear transformation $-[C + D_1]^{-1}C$ to this estimate.

The results of this paper can also be extended to include the limiting case where P, which is defined by (13b), is nonnegative definite. Another limiting case, corresponding to $D_1 = \underline{0}$, is related to a particular minimax estimation problem with a generalized quadratic loss function; here one is seeking an estimate for a guassian random vector with unknown mean but known covariance, which is observed in additive guassian noise. Results for this minimax estimation problem appear in [1].

The stochastic game discussed in this paper can be viewed as a special case of discrete stochastic pursuit-evasion game of the linear quadratic guassian (LQG) variety where one player (the pursuer) makes noisy state measurements and the other player (the evader) makes no state measurements. A general pursuit-evasion game of this variety, considered as a problem in the continuous time domain, is discussed in [6]. Reference [6] states that the evader's maximin policy depends only on the system dynamics and prior statistical data where all of this information is known to both players. In particular, when the intial state x(0) of the dynamical system is a zero mean normal random vector, then the maximin stratgey for the evader is the control function v(.) defined by $v(t) = \underline{o}$, for all t in [0,T]. This policy is a valid maximin strategy provided that the solution to a certain matrix Riccati equation exists for all t in [0,T]. (The precise Riccati equation is not explicitly stated in this reference. However, it can be obtained from a standard control theoretic argument.) The corresponding minimax control policy for the pursuer is determined by the solution to a certain LQG optimal control problem. At this point the following question remains:

What can be said about the solution to this pursuit-evasion game when the solution to the relevant Riccati equation fails to exist for all t in the closed interval $[0,T]$?

This question can be answered as follows, when this pursuit evasion problem is couched in the discrete time domain. (The analogous solution for the continuous time domain problem will be considered in a forthcoming paper.) For the general case, the existence condition involving the Riccati equation is replaced by conditions on the eigenvalues of a certain symmetric matrix, which we shall denote by ν in the following discussion. For the specific game which we consider in this paper, this matrix ν is denoted by M and is defined by (15). The condition that the matrix ν be nonpositive definite is analogous to the condition that the solution to the Riccati equation exist on the interval $[0,T]$. The discrete time maximin control policy for the evader, which is analogous to the continuous time maximin policy $\{v(t) = \underline{o};\ \ t \in [0,T]\}$, is the control policy defined by the "zero sequence" of control vectors, that is, the sequence $\{v(i) = \underline{o};\ \ i = 0,1,\ldots,N-1\}$. For the specific game considered in this paper, the condition $M \leq \underline{0}$ insures that $v^* = \underline{o}$ is a maximin policy for player 2. If the symmetric matrix ν fails to be nonpositive definite, then a maximin policy for the evader is a zero mean gaussian random process, with a covariance function which is determined by a solution to a nonlinear matrix equation. For the specific game considered in this paper, this reduces to the cases where M fails to be nonpositive definite. In this circumstance, a maximin policy for player 2 is $v^* \in N[\underline{o},\Lambda^*]$, where Λ^* satisfies (6a) and (6b). Note that the "zero solution" can be regarded as a special case of the more general gaussian result. The details of the analysis of the general discrete LQG pursuit-evasion game will be reported elsewhere.

In summary, the stochastic game discussed in this paper can be viewed as a special case of a discrete stochastic pursuit-evasion game of the LQG variety where one player (the pursuer) makes noisy state measurements and the other player (the evader) makes no state measurements. In this regard, the major contribution of this paper is a rigorous proof of the existence of a maximin policy, defined by a gaussian measure, for the evader in this specific game; this proof can be extended to include the general LQG pursuit-evasion game. The existence of a maximin policy defined by a gaussian measure, provides new insight into the existence, structure and behavior of the solution of these stochastic games.

APPENDIX I

The following observations and lemmas are used to prove part (i) of the Theorem.

Observation 1

Let Ω denote the space of all n x n real symmetric matrices. Further define a metric $\rho(W,M):\ \Omega \times \Omega \to E^1$ by

$$\rho(W,M) = \left[\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \leq j}}^{n} (W_{ij} - M_{ij})^2 \right]^{1/2} . \tag{1-1}$$

With these definitions, the metric space (Ω,ρ) is isomorphic to $E^{\frac{n(n+1)}{2}}$.

Observation 2

Define sets T_1 and $T_2(\alpha)$ as follows:

$$T_1 \triangleq \{\Lambda \in \Omega :\ \Lambda \geq \underline{0}\ \ , \tag{1-2a}$$

$$T_2(\alpha) \triangleq \{\Lambda \in \Omega : \rho(\Lambda,\underline{0}) \le \alpha\} \quad . \tag{1-2b}$$

With these definitions, the set $S(\alpha)$ defined by

$$S(\alpha) \triangleq T_1 \cap T_2(\alpha) = \{\Lambda \in \Omega : \Lambda \ge \underline{0}, \rho(\Lambda,\underline{0}) \le \alpha\} \quad , \tag{1-3}$$

is both compact and convex in (Ω,ρ).

Lemma 3

Consider the following abstract game:

Define $J(\Lambda_1,\Lambda_2)$ by

$$\begin{aligned} J(\Lambda_1,\Lambda_2) = {} & \mathrm{Tr}[K'(\Lambda_1)[C + D_1]K(\Lambda_1)R] \\ & + \mathrm{Tr}[\{[I + K(\Lambda_1)]'C[I + K(\Lambda_1)] + K'(\Lambda_1)D_1K(\Lambda_1)\}Q] \\ & + \mathrm{Tr}[\{[I + K(\Lambda_1)]'C[I + K(\Lambda_1)] + K'(\Lambda_1)D_1K(\Lambda_1) - D_2\}\Lambda_2] \quad , \end{aligned} \tag{1-4}$$

where

$$K(\Lambda_1) \triangleq -[C + D_1]^{-1}C[\Lambda_1 + Q][\Lambda_1 + Q + R]^{-1} \quad . \tag{1-5}$$

Then,

$$\min_{\Lambda_1 \in S(\alpha)} \; \max_{\Lambda_2 \in S(\alpha)} J(\Lambda_1,\Lambda_2) = \max_{\Lambda_2 \in S(\alpha)} \; \min_{\Lambda_1 \in S(\alpha)} J(\Lambda_1,\Lambda_2)$$

Proof:

The proof of this lemma is in two parts:

Part (I):

Since the kernel $J(\Lambda_1,\Lambda_2)$ is continuous in the pair (Λ_1,Λ_2), and since $S(\alpha)$ is compact, there exist two probability measures on $S(\alpha)$ denoted by μ_1^*, μ_2^* which are optimal mixed strategies for this game. [See reference [5], p. 88].

Part (II):

It will be demonstrated in what follows that the probability measures μ_1^* and μ_2^* referred to in Part (I) are one-point; that is, the game has a solution in pure strategies.

Observation 3

Let Λ_2 be any fixed element of T_1 which is defined by (1-2a). Then $J(\Lambda_1,\Lambda_2)$ is minimized when $\Lambda_1 = \Lambda_2$.

Proof:

The proof follows by identifying this minimization problem with a standard stochastic minimization problem. Let the observation z be given by

$$z = x + v + w \quad , \tag{1-6a}$$

where

$$x \in N[\underline{o},Q], \; v \in N[\underline{o},\Lambda_2], \; w \in N[\underline{o},R]$$

and the random vectors x, v, and w are statistically independent. Further, let a quadratic payoff function $L_1(u(z))$ be defined by

$$L_1(u(z)) = [u + x + v]'C[u + x + v] + u'D_1u - v'D_2v \quad , \tag{1-6b}$$

where $C \ge \underline{0}$, $D_1 > \underline{0}$, $D_2 > \underline{0}$, and $u(.) \in \Sigma$, where Σ is the space of all Borel measurable functions mapping E^n into E^n.

Then, the policy $u(.)$ that minimizes the average payoff $E[L_1(u(z))]$, where the expectation operation is taken over the prior statistics of x, v, and w, is given by

$$u(z) = -[C + D_1]^{-1}CE[x + v \,|z] \quad . \tag{1-7a}$$

This result follows from a standard statistical argument which will be omitted here. Since $E[x + v\,|z] = [\Lambda_2 + Q][\Lambda_2 + Q + R]^{-1}z$, expression (1-7a) is equivalent to

$$u(z) = K_1z \quad , \tag{1-7b}$$

where K_1 is given by

$$K_1 \triangleq -[C + D_1]^{-1}C[\Lambda_2 + Q][\Lambda_2 + Q + R]^{-1} \quad . \qquad (1\text{-}7c)$$

The average payoff using u(z) given by (1-7b) is equal to $J(\Lambda_2,\Lambda_2)$. Comparing K_1 with $K(\Lambda_1)$ defined by (1-5), it can then be concluded that $J(\Lambda_1,\Lambda_2)$ is minimized for fixed $\Lambda_2 \in T_1$ when $\Lambda_1 = \Lambda_2$.

Now consider the lower value of the game over (μ_1,μ_2). That is, first fix μ_2 and then minimize with respect to μ_1. Further let

$$\tilde{J}(\Lambda_1|\mu_2) \triangleq E_{\mu_2}[J(\Lambda_1,\Lambda_2)] \quad . \qquad (1\text{-}8)$$

("E_{μ_i} [.]" is "expected value of [.] with respect to measure μ_i".)
Since $J(\Lambda_1,\Lambda_2)$ defined by (1-4) is linear in Λ_2,

$$\tilde{J}(\Lambda_1|\mu_2) = J(\Lambda_1,E_{\mu_2}[\Lambda_2]) \quad . \qquad (1\text{-}9)$$

A measure μ_1 that minimizes $E_{\mu_1}[\tilde{J}(\Lambda_1|\mu_2)]$ is the measure that assigns probability one to the element $\Lambda_1 = E_{\mu_2}[\Lambda_2]$. [This follows from Observation 3]. Hence for any μ_2, the related minimizing measure μ_1 is one-point. To complete the evaluation of the lower value of the game with respect to the mixed strategies (μ_1,μ_2), it is sufficient to compute

$$J_\ell \triangleq \max_{\mu_2} J(E_{\mu_2}[\Lambda_2], E_{\mu_2}[\Lambda_2]) \quad . \qquad (1\text{-}10)$$

Observations:

(a) Since μ_2 is a probability measure on $S(\alpha)$, and since $S(\alpha)$ is convex, $E_{\mu_2}[\Lambda_2] \in S(\alpha)$.

(b) The $\max\limits_{\Lambda \in S(\alpha)} J(\Lambda,\Lambda)$ exists since $J(\Lambda,\Lambda)$ is continuous in Λ and $S(\alpha)$ is compact.

Let Λ^* maximize $J(\Lambda,\Lambda)$ over $S(\alpha)$. A maximizing measure μ_2 defined by (1-10) is given by the measure which assigns probability one to the element Λ^*. Hence there exists a one-point measure $\underline{\mu}_2$ associated with the lower value of the game.

Let $(\bar{\mu}_1,\mu_2)$ denote the measures associated with the upper value of the game.

From Part (I) it follows that $\underline{\mu}_1 = \bar{\mu}_1$, and $\underline{\mu}_2 = \bar{\mu}_2$. But since $(\underline{\mu}_1,\underline{\mu}_2)$ are one-point, this completes the proof of the lemma.

In addition to proving Lemma 3, it has been shown that the solution to the game is given by $\Lambda_1 = \Lambda_2 = \Lambda^*$, where Λ^* maximizes $J(\Lambda,\Lambda)$ over $S(\alpha)$.

Discussion on Lemma 3

The solution $(\Lambda_1 = \Lambda^*(\alpha),\ \Lambda_2 = \Lambda^*(\alpha))$ to the game defined in Lemma 3 will depend in general on the value α which defines $S(\alpha)$. In what follows it will be shown that the solution $(\Lambda^*(\alpha),\ \Lambda^*(\alpha))$ will be independent of α for sufficiently large α, if

$$C[C + D_1]^{-1}D_1 - D_2 < \underline{0} \quad . \qquad (1\text{-}11)$$

Hence it will be possible to replace $S(\alpha)$ with T_1 and insure the existence of a solution to the game in pure strategies over T_1, whenever condition (1-11) is satisfied.

When $\Lambda_1 = \Lambda_2 = \Lambda$, and the expression for $K(\Lambda_1)$ given by (1-5) is substituted into (1-4), the expression for $J(\Lambda_1,\Lambda_2)$ simplifies to:

$$\begin{aligned} J(\Lambda,\Lambda) &= \mathrm{Tr}[C(\Lambda + Q)] \\ &\quad - \mathrm{Tr}[(\Lambda + Q)C(C + D_1)^{-1}C(\Lambda + Q)(\Lambda + Q + R)^{-1}] - \mathrm{Tr}[D_2\Lambda] . \end{aligned} \qquad (1\text{-}12a)$$

Note that $J(\underline{0},\underline{0})$ is given by

$$J(\underline{0},\underline{0}) = \mathrm{Tr}[CQ] - \mathrm{Tr}[QC(C + D_1)^{-1}CQ(Q + R)^{-1}] \quad . \tag{1-12b}$$

Hence, using the expression for $J(\underline{0},\underline{0})$ and simplifying, expression (1-12a) becomes

$$\begin{aligned} J(\Lambda,\Lambda) = {} & J(\underline{0},\underline{0}) - \mathrm{Tr}[C(C + D_1)^{-1}CR(Q + R)^{-1}Q] + \mathrm{Tr}[C(C + D_1)^{-1}CR] \\ & + \mathrm{Tr}[(C[C + D_1]^{-1}D_1 - D_2)\Lambda] - \mathrm{Tr}[C[C + D_1]^{-1}CR[\Lambda + Q + R]^{-1}R] \end{aligned} \tag{1-12c}$$

Since $\mathrm{Tr}[C(C + D_1)^{-1}C[R(Q + R)^{-1}Q + R(\Lambda + Q + R)^{-1}R]] \geq \underline{0}$,

$$J(\Lambda,\Lambda) \leq J(\underline{0},\underline{0}) + \mathrm{Tr}[C(C + D_1)^{-1}CR] - \mathrm{Tr}[(D_2 - C[C + D_1]^{-1}D_1)\Lambda] \tag{1-13}$$

for all $\Lambda \in T_1$. It follows from inequality (1-13) that there exists a finite real number $\eta \geq 0$ such that $||\Lambda|| > \eta$ implies $J(\Lambda,\Lambda) < J(\underline{0},\underline{0})$. A suitable choice of η is

$$\eta^* = \mathrm{Tr}[C(C + D_1)^{-1}CR]/[D_2 - C(C + D_1)^{-1}D_1]_{\min} \quad , \tag{1-14}$$

where $[.]_{\min}$ denotes the minimum eigenvalue of $[.] > \underline{0}$.

It follows from (1-13) that if $\alpha > \eta^*$, then $\Lambda^*(\alpha) \in S(\eta^*)$. Thus, if the condition given by (1-11) holds, then $\Lambda^*(\alpha)$ becomes independent of α when $\alpha > \eta^*$. Hence, if $[C[C + D_1]^{-1}D_1 - D_2] < \underline{0}$, then $S(\alpha)$ can be replaced by T_1 in the hypothesis of Lemma 3.

Lemma 4

The matrix $\Gamma(\Lambda^*)$ defined by (5a) and (5b) is nonpositive definite, where Λ^* denotes a solution to the game defined in Lemma 3 with $S(\alpha)$ replaced by T_1 and with condition (1-11) satisfied.

Proof by contradiction: Write $J(\Lambda_1,\Lambda_2)$ as

$$J(\Lambda_1,\Lambda_2) = \mathrm{Tr}[K'(C + D_1)KR] + \mathrm{Tr}[(\Gamma(\Lambda_1) + D_2)Q] + \mathrm{Tr}[\Gamma(\Lambda_1)\Lambda_2], \tag{1-15}$$

with K defined by (1-5) and $\Gamma(\Lambda_1)$ defined by (5a). Let Λ_1^* be an optimal choice for Λ_1 and further assume that $\Gamma(\Lambda_1^*)$ has at least one positive eigenvalue. Then it is possible to find a $\Lambda_2^* \in T_1$ that will make $J(\Lambda_1^*,\Lambda_2)$ arbitrarily large. But this is a contradiction since the upper value of the game is bounded when condition (1-11) is satisfied. Therefore $\Gamma(\Lambda_1^*)$ cannot have any positive eigenvalues.

Lemma 5

Let Λ^* denote a solution to the game defined in Lemma 3 with $S(\alpha)$ replaced by T_1, and with condition (1-11) satisfied. Then,

$$\Gamma(\Lambda^*)\Lambda^* = \underline{0} \quad .$$

Proof: $J(\Lambda_1,\Lambda_2)$ is defined by (1-15). Let Λ_1^* denote an optimal choice of Λ_1, and Λ_2^* denote an optimal choice of Λ_2. Assume $\Gamma(\Lambda_1^*)\Lambda_2^* \neq \underline{0}$. But $\Gamma(\Lambda_1^*) \leq \underline{0}$ by Lemma 4, and further $\Lambda_2^* \geq \underline{0}$ by definition. Note that $\mathrm{Tr}[AB] = 0$ if and only if $AB = \underline{0}$ when $A \leq \underline{0}$ and $B \geq \underline{0}$; hence it follows that $\mathrm{Tr}[\Gamma(\Lambda_1^*)\Lambda_2^*] < 0$. This contradicts the optimality of Λ_2^* since then $J(\Lambda_1^*,\underline{0}) > J(\Lambda_1^*,\Lambda_2^*)$. Hence, $\Gamma(\Lambda_1^*)\Lambda_2^* = \underline{0}$. But $\Lambda_1^* = \Lambda_2^* = \Lambda^*$ by Lemma 3. Therefore $\Gamma(\Lambda^*)\Lambda^* = \underline{0}$.

Proof of part (i) of the Theorem.

Assume that the condition given by (7) holds. If $\Gamma(\Lambda)$ is defined by (5a), (5b), and (5c), then there exists at least one symmetric matrix $\Lambda^* \geq \underline{0}$ which satisfies (6a) and (6b). Lemmas 4 and 5 prove this part of the Theorem by making use of the abstract game defined in Lemma 3 with $S(\alpha)$ replaced by T_1.

REFERENCES

[1] T. Basar and M. Mintz, "On Linear Filters and Minimax Estimation", Proceedings of The Fifth Princeton Conference On Information Sciences And Systems, March, 1971.

[2] R. Bellman, "Some Inequalities For The Square Root Of A Positive Definite Matrix", Linear Algebra And Its Applications, I, (1968), 321-324.

[3] T. S. Ferguson, Mathematical Statistics, A Decision Theoretic Approach, Academic Press, New York, 1967.

[4] E. L. Lehmann, Notes On The Theory Of Estimation, University of California Press, Berkeley, California, 1966, Ch. 1-5.

[5] G. Owen, Game Theory, Saunders, Philadephia, 1968.

[6] I. B. Rhodes and D. G. Luenberger, "Differential Games With Imperfect State Information", IEEE Transactions On Automatic Control, Vol. AC-14, No. 1, (1969), 29-38.

[7] D. Sworder, Optimal Adaptive Control Systems, Academic Press, New York, 1966.

[8] A. Wald, Statistical Decision Functions, Wiley, New York, 1950.

NOTATION

The following notation applies unless otherwise noted:

Lower case letters e. g. v, z, denote real vectors of dimension n.

Upper case letters e. g. R, Q, denote real $n \times n$ matrices.

Ω denotes the space of all $n \times n$ real symmetric matrices.

Σ denotes the space of all Borel measurable functions mapping E^n into E^n.

$\underline{0}$ denotes the $n \times n$ zero matrix.

$\underline{o}$ denotes the n dimensional zero vector.

I denotes the $n \times n$ identity matrix.

v' denotes the transpose of v.

$v \in N[\underline{o},\Lambda]$ denotes the statement that v is a normal random vector with zero mean and covariance Λ.

$\lambda = N[\underline{o},\Lambda]$ denotes the statement that the probability measure λ is normal with zero mean and covariance Λ.

$A > \underline{0}$; A is a symmetric positive definite matrix.

$A \geq \underline{0}$; A is a symmetric nonnegative definite matrix.

$\mathrm{Tr}[A]$ denotes the trace of the matrix A.

A_{ij} denotes the ij'th element of the matrix A.

$A^{1/2}$ denotes the unique symmetric nonnegative definite square root of the matrix A.

LINEAR DIFFERENCE GAMES WITH DELAYS

JERRY L. HANSON, PhD
Lt Colonel, USAF
Department of Electrical Engineering
Air Force Institute of Technology
Wright-Patterson AFB, Ohio 45433

ABSTRACT

The problem considered is the optimal solution of the sampled-data version of a linear differential game with constant time delays in the state, minimizing control and/or maximizing control and with a quadratic payoff. The foundation for the problem solution is in the areas of optimal control, time-lag systems and differential games.

The discrete state equation for the resulting difference game is a finite approximation to the infinite series which exactly describes the game at the sampling instants. One method of achieving the finite, discrete state equation utilizes the state transition matrix and a "straight-line" approximation of the state trajectory over past sampling periods.

A "history state" vector is defined which transforms the game into a game with no delays. Dynamic programming is then used to synthesize the optimal game.

A first-order example is presented to illustrate the method, the effect of the choice of sampling period on the game and the effect of a delay in the state on the game.

I. INTRODUCTION

Time delays are inherent in many systems and may often be neglected in system modeling because of the mathematical complications which they add to the problem. Significant analytical and numerical results for time-delay systems are difficult to find in the literature. This is primarily because of the computational difficulties implied by the existence and uniqueness theorems which have been derived for the time-delay optimal control problems, including those which are continuous[1-8] or discrete[9,10] in nature. This is particularly true for differential games with delays.[2,11,12]

The objective of this paper is to present a method of synthesis for a class of linear differential games with delays. The method is direct and computationally feasible and could lead to significant analytical and numerical results. Although this is the specific problem considered, it is also hoped that interest may be revived in several ideas previously introduced for optimal control problems with delays.

The method is basically an application and extension of three basic concepts. The first is the use of a "straight-line" approximation of the state vector over past sampling periods connecting known or computed points on the state trajectory.[13] This is done to obtain a finite approximation of the infinite series[14] which accurately describes the state at the sampling times for the game with piecewise-constant controls.

The finite difference game with delays is transformed into a difference game without delays by using the concept of an enlarged vector. This technique was first suggested for optimal control problems with delays in 1960[15] and again in 1968.[13] The chief objection raised [9,16] to this technique is the increased dimensionality of the resulting problem. This objection is outweighed by the directness of the method and by taking advantage of the inherent "emptiness" (many o's) of the resulting matrices in the computational scheme.

Finally, dynamic programming and the principle of optimality[17] are used to synthesize the optimal game, i.e. the optimal strategies, state trajectory and value of the game, in a manner similar to that applied to difference games without delays.[18]

II. A LINEAR DIFFERENTIAL GAME WITH DELAYS

Consider a system described by:

$$\dot{z}(t) = A_o z(t) + A_1\, z(t-d) + B_1\, u(t-\ell) + C_1 w(t-r) \qquad (1)$$

for t in (t_o, t_f), where z(t) is an nx1 state vector and where u(t) and w(t) are scalar controls. (Note: The addition of multiple delays and vector controls requires no new techniques and thus are not included here in the interest of space and time. These cases and the time-varying case are treated in detail in reference 19.)

$z_o(t)$, $u_o(t)$ and $w_o(t)$ are known, continuous functions such that:

$$z(t) = z_o(t) \text{ on } (t_o-d,\ t_o)$$

$$u(t) = u_o(t) \text{ on } (t_o-\ell,\ t_o) \qquad (2)$$

and $$w(t) = w_o(t) \text{ on } (t_o-r,\ t_o)$$

A quadratic payoff is defined by:

$$P_f(u,w) = \int_{t_o}^{t_f} [z'(t)Q(t)z(t) + R(t)u^2(t) - S(t)w^2(t)]dt \qquad (3)$$

where Q(t) is an nxn continuous, symmetric, positive semi-definite matrix; $R(t) > o$ and $S(t) > o$. The object of u(t) is to minimize P_f. The object of w(t) is to maximize P_f.

III. THE LINEAR DIFFERENCE GAME WITH DELAYS

Consider the sampled-data version of the differential game with piecewise constant controls, i.e. let

$$u(t) = u(kT) \text{ and } w(t) = w(kT) \text{ for } kT \leq t < (k+1)T \tag{4}$$

and

$$t_f - t_o = NT,\ d = DT,\ \ell = LT \text{ and } r = RT \tag{5}$$

A solution to Eq. (1) can be expressed as

$$z(t) = \phi(t-t_o)z(t_o) + \int_{t_o}^{t} \phi(t-s)[A_1 z(s-d) + B_1 u(s-\ell) + C_1 w(s-r)]ds \tag{6}$$

where

$$\phi(t) = e^{A_o t} \tag{7}$$

For each sampling period in the past, assume a "straight-line" approximation of the state connecting the end points, i.e.

$$z(s-d) = z(s-DT) \approx z[(k-D)T] + [z[(k-D+1)T] - z[(k-D)T]]\frac{s-kT}{T} \tag{8}$$

For $kT \leq s \leq (k+1)T$. Now substitute $t_o = kT$, $t = (k+1)T$, Eqs. (4), (5) and (8) into Eq. (6). Dropping the T from the notation (i.e. $k \rightarrow kT$, etc.) the resulting difference equation can be written as

$$z(k+1) = E_o z(k) + E_1\, z(k-D+1) + F_1\, z(k-D) + G_1 u(k-L) + H_1 w(k-R) \tag{9}$$

If

$$\phi_1(T) = \int_o^T \phi(s)ds \tag{10}$$

and

$$\phi_2(T) = \frac{1}{T}\int_o^T s\phi(s)ds \tag{11}$$

where $\phi(t)$ is given by (7), then

$$E_o = \phi(T),\ E_1 = [\phi_1(T) - \phi_2(T)]A_1,\ F_1 = \phi_2(T)A_1,$$
$$G_1 = \phi_1(T)B_1 \text{ and } H_1 = \phi_1(T)C_1 \tag{12}$$

The approximation given by (8) eliminates the infinite series[14] which results from the first term under the integral in Eq. (6). Therefore, the sampling period T should not be chosen indiscriminately. One should consider the smoothness of the initial trajectory $z_o(t)$ as well as the value of the time constants and natural frequencies of the system with zero delay in the state.

To complete the game definition, a quadratic-sum payoff corresponding to Eq. (3) will be given by

$$P_N(u,w) = \sum_{k=1}^{N} T[z'(k)Q(k)z(k) + R(k-1)u^2(k-1) - S(k-1)w^2(k-1)] \tag{13}$$

If a pure optimal strategy (u^o, w^o) exists, the value of the resulting game is defined as

$$V_N = P_N(u^o,w^o) = \min_u \max_w P_N(u,w) = \max_w \min_u P_N(u,w) \qquad (14)$$

IV. THE ENLARGED OR "HISTORY" SYSTEM

A vector h(k) will be defined and called the "history vector" because it will include both past and present information about the system. In terms of h(k) Eq. (9) can be expressed as

$$h(k+1) = A\,h(k) + B\,u(k) + C\,w(k) \qquad (15)$$

where [for the case when $D > 0$, $L > 0$ and $R = 0$]

$$h(k) = \left[\begin{array}{c} z(k) \\ z(k-1) \\ \cdot \\ \cdot \\ \underline{z(k-D)} \\ u(k-1) \\ \cdot \\ \cdot \\ u(k-L) \end{array}\right], \quad B = \left[\begin{array}{c} 0 \\ \cdot \\ \cdot \\ \cdot \\ \underline{0} \\ 1 \\ 0 \\ \cdot \\ 0 \end{array}\right] \qquad C = \left[\begin{array}{c} H_1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{array}\right]$$

and

$$A = \left[\begin{array}{cccccc|ccccc}
E_o & 0 & \cdot & 0 & E_1 & F_1 & 0 & \cdot & \cdot & 0 & G_1 \\
1 & & & & & & & & & & \\
 & \ddots & & 0 & & & & & & 0 & \\
 & 0 & \ddots & & & & & & & & \\
 & & & \ddots\,1 & 0 \cdot & 0 & & & & & \\ \hline
 & & & & & & 0 & & & & \\
 & & 0 & & & & 1 & & & 0 & \\
 & & & & & & & \ddots & & & \\
 & & & & & & & 0 & \ddots\,1 & 0 &
\end{array}\right] \qquad (16)$$

where all blank spaces indicate 0's. Note that there are eight possible cases resulting from different combinations of $D \geq 0$, $L \geq 0$ and $R \geq 0$.

Noting from (16) that

$$z(k) = [I\!: 0 \ldots 0]h(k) \qquad (17)$$

the payoff in Eq. (13) can be rewritten as

$$P_N[h(0),\, u,w] = \sum_{k=1}^{N} [h'(k)H(k)h(k) + U(k-1)u^2(k-1) - W(k-1)w^2(k-1)] \qquad (18)$$

where

$$H(k) = T\left[\begin{array}{c|c} Q(k) & 0 \\ \hline 0 & 0 \end{array}\right], \quad U(k-1) = TR(k-1) \text{ and } W(k-1) = TS(k-1) \qquad (19)$$

The initial history vector, h(0), is known from Eqs. (2) and (16).

V. THE OPTIMAL GAME

The optimal game is synthesized by considering an (N-j) stage game starting at k = j+1, i.e. the last (N-j) stages of the original N stage game. Eq. (18) and the principle of optimality result in

$$V_{N-j}[h(j)] = \min_{u(j)} \max_{w(j)} [h'(j+1)\ H(j+1)\ h(j+1) + U(j)u^2(j) - W(j)w^2(j)] + V_{N-j-1}[h(j+1)] \tag{20}$$

Based on the results of the one-stage game,[19] the value is assumed (and later proved) to be quadratic in h, i.e.

$$V_{N-j-1}[h(j+1)] = h'(j+1)K(N-j-1)h(j+1) \tag{21}$$

where K is symmetric and must also be positive-definite. Using Eqs. (15) and (21) and symmetry, Eq. (20) can be expressed as

$$P_{N-j} = h'M_{AA}h + (U+M_{BB})u^2 - (W-M_{CC})w^2 + 2(uM_{BA}+wM_{CA})h + 2M_{BC}uw \tag{22}$$

where the arguments have been omitted to simplify notation. if

$$L(N-j-1) = H(j+1) + K(N-j-1) \tag{23}$$

then

$$M_{XY} = M'_{YX} = X'L(N-j-1)Y \tag{24}$$

Applying the necessary conditions for optimality, i.e.

$$\frac{\partial P_{N-j}}{\partial u} = 0 \text{ and } \frac{\partial P_{N-j}}{\partial w} = 0 \tag{25}$$

to Eq. (22) and adding the arguments again leads to

$$u^o(j) = -F(N-j)h(j) \tag{26}$$

and

$$w^o(j) = -G(N-j)h(j) \tag{27}$$

where

$$F(N-j) = \{M_{BC}(N-j-1)M_{CA}(N-j-1) + [W(j)-M_{CC}(N-j-1)]M_{BA}(N-j-1)\}/DEN, \tag{28}$$

and

$$G(N-j) = \{M_{BC}(N-j-1)M_{BA}(N-j-1) - [U(j)+M_{BB}(N-j-1)]M_{CA}(N-j-1)\}/DEN \tag{29}$$

with

$$DEN = [U(j)+M_{BB}(N-j-1)][W(j)-M_{CC}(N-j-1)] + M^2_{BC}(N-j-1) \tag{30}$$

Substituting Eqs. (26) and (27) into (22) leads to the assumption proposed in (21) where if

$$D(N-j) = A - BF(N-j) - CG(N-j) \tag{31}$$

then

$$K(N-j) = D'(N-j)L(N-j-1)D(N-j)+F'(N-j)U(j)F(N-j)-G'(N-j)W(j)G(N-j) \tag{32}$$

Starting with K(0) = 0, Eqs. (23), (24), and (28) - (32) can be used recursively to compute the F, G, D and K matrices with arguments 1 → N.

Starting with h(0), Eq. (15) which can now be written as

$$h(k+1) = D(N-k)h(k) \tag{33}$$

can be used to compute h(1) → h(N).

Finally, the value of the game can be computed from Eq. (21) as

$$V_N[h(0)] = h'(0)K(N)h(0) \tag{34}$$

Note that if the H, U and W matrices are constant, Eq. (34) can be used at each stage of the computation to compute the value of any length game up to and including N.

Sufficient conditions for a pure optimal strategy, i.e.

$$\frac{\partial^2 P_{N-j}}{\partial u^2} > 0 \text{ and } \frac{\partial^2 P_{N-j}}{\partial w^2} < 0 \tag{35}$$

when applied to Eq. (22) lead to

$$U(j) + M_{BB}(N-j-1) > 0 \tag{36}$$

which is always true from the original problem specifications, and

$$W(j) - M_{CC}(N-j-1) > 0 \quad . \tag{37}$$

Condition (37) must be checked at each step of the game. If condition (37) does not hold true, a unique, finite optimal strategy does not exist, i.e. a saddle point does not exist and the game cannot proceed in an optimal manner.

VI. AN EXAMPLE

Consider a first-order game described by

$$\dot{z}(t) = 0.5\ z(t-d) + 0.01\ u(t) + 0.01\ w(t) \tag{38}$$

with

$$u(t) = w(t) = 0 \text{ for } t < 0 \tag{39}$$

and

$$z(t) = 1 \text{ for } t \leq 0 \tag{40}$$

With d = 0, the sampled-data game with piecewise constant controls can be described at the sampling times by

$$z(k+1) = (e^{.5T})\ z(k) + .02(e^{.5T}-1)[u(k) + w(k)] \tag{41}$$

with the use of Eqs. (10) - (12). Note that this is already in the form of Eq. (15). With d = 1, the use of the approximation in Eq. (8) and thence Eqs. (10) - (12) lead to

$$z(k+1) \simeq z(k) + {}^{T}/4\ z(k-{}^{1}/T+1) + {}^{T}/4\ z(k-{}^{1}/T) + 0.01T[u(k)+w(k)]. \tag{42}$$

Recall from Eqs. (6) and (8) that the area under the state curve over a past sampling period has been approximated by the area under a straight-line connecting the end-points. Contrast this with the common approximation

$$\dot{z}(t) \simeq \dot{z}(kT) \text{ for } kT \leq t < (k+1)T \tag{43}$$

and

$$z[(k+1)T] \approx z(kT) + T\,\dot{z}(kT) \tag{44}$$

which when applied to (38) yields

$$z(k+1) \approx z(k) + {}^{T}/2\; z(k-{}^{1}/T) + 0.01\,T[u(k) + w(k)] \tag{45}$$

which is the same as assuming a constant state over the past sampling period.

Let the payoff for the game be given by

$$P_N = \sum_{k=1}^{N} T[(kT)z^2(k) + u^2(k-1) - 2w^2(k-1)] \tag{46}$$

with

$$N = {}^{5}/T \quad (\text{i.e. } t_f = 5 \text{ seconds}) \tag{47}$$

h, A, B, C, H, U and W for Eqs. (15) - (19) can then be defined for any value of T.

$T = 1 = {}^{1}/2$ of the time constant for the system with no delay seems like a reasonable choice for the sampling period. However, games with no delay (nd) and with a 1-second delay (d) were solved with T = 1, 1/2 and 1/4.[19] In each case the players started the game with maximum effort [$u\ (\max) < 0$ and $w\ (\max) > 0$] and tapered off while the state started at $z = 1$ and increased toward $z\ (\max)$ at the end of the game. The value of the game [because of (46)] and the optimal strategies [because of more choices] varied somewhat with the sampling period T. However, the state trajectory and the effect of the delay on the game (as shown below) were effected very little by the choice of T.

Delay/No Delay	T = 1	T = 1/2	T = 1/4
$V_d \,/\, V_{nd}$	.303	.313	.321
$u_d(\max)/u_{nd}(\max)$	.230	.229	.230
$w_d(\max)/w_{nd}(\max)$	.230	.229	.230
$z_d(\max)/z_{nd}(\max)$	.526	.519	.515

Thus, the increases in the dimensionality of (15) - (19) and in computational time caused by the smaller values of T are not justified by the results.

Note that in this example, the state-delay worked to the advantage of the minimizing player and had a significant effect on the game.

VII. CONCLUSIONS

Optimal strategies for the sampled-data version of a class of linear differential games with piecewise-constant controls and quadratic payoff have been considered. The method of solution results in a reasonable (with the proper choice of sampling period), finite approximation to a problem which is essentially infinite in dimension.

Computationally, the method is direct and can be extended to cases with vector controls and multiple delays. A subsequent paper will discuss how the dimensions of the "history state" matrices can be effectively reduced in the computational scheme.

A simple first order example illustrated two points. The increase in the dimension of the problem and in the required computational time which results from reducing the sampling period is not justified, beyond a certain point, by the results. A delay in the state can have a significant effect on the game.

A subsequent paper will present some results on an extensive study of the effect of delays in the state, minimizing control and/or maximizing control on three second-order games (stable, conditionally stable and unstable) previously solved for games without delays.[18]

REFERENCES

1. Pontryagin, L.S., Boltyanski, V.G., Gamkrelidze, R.V., and Mischenki, E.F., The Mathematical Theory of Optimal Processes, Interscience Publishers, New York, 1962.

2. Oguztoreli, M.N., Time-Lag Control Systems, Academic Press, New York, 1966.

3. Nosov, V.R., "On a Problem Occurring in the Theory of Optimum Control with After-effects," PMM J. Of Applied Mathematics and Mechanics, 30, 2, 1966, pp. 481-486.

4. Kharatishvili, G.L., "A Maximum Principle in Extremal Problems with Delays," Mathematical Theory of Control (A.V. Balakrishman and L.W. Neustadt, Eds.), Academic Press, New York, 1967.

5. Popov, V.M., and Khalanay, A., "A Problem in the Theory of Time-Delay Optimum Systems," Automation and Remote Control, 24, 2, February, 1963, pp. 129-131.

6. Chyung, D.H., and Lee, E.B., "Linear Optimal Systems with Time Delays," J. SIAM on Control, 4, 3, 1966, pp. 548-575.

7. Chyung, D.H., "Linear Time Lag Systems with Side Constraints," IEEE Trans. on Automatic Control, AC-12, 4, August, 1967, pp. 613-615.

8. Chyung, D.H., "Discrete System with Delays in Control," IEEE Trans. on Automatic Control, AC-14, 2, April 1969, pp. 196-197.

9. Chyung, D.H., "Discrete Optimal Systems with Time Delays," IEEE Trans. on Automatic Control, AC-13, February, 1968, p. 117.

10. Chyung, D.H., "Discrete Systems with Delays in Control," Proc. Sixth Annual Allerton Conference on Circuit and System Theory, 1968, pp. 465-467.

11. Oguztoreli, M.N., "Optimal Pursuit Strategy with Retarded Control Systems," J.SIAM on Control, 2, 1, 1964, pp. 89-105.

12. Halanay, A., "Differential Games with Delay," J. SIAM on Control, 6, 4, 1968, pp. 579-593.

13. Day, K.S., "Optimal Control of Linear Time Lag Systems," Proc. JACC, 1968, pp. 1046-1053.

14. Koepcke, R.W., "On the Control of Linear Systems with Pure Time Delay," Proc JACC, 1964, pp. 397-403.

15. Kramer, J.D.R., Jr., "On Control of Linear Systems with Time Lags," Information and Control, 3, 4, 1960, pp. 299-326.

16. McKinnon, Duncan, "Optimal Control of Systems with Pure Time Delays Using A Variational Programming Approach," IEEE Trans. on Automatic Control, AC-12, 3, 1967, pp. 255-262.

17. Bellman, R.E., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

18. Shea, P.D., "Optimal Control of Linear Differential and Difference Games," PhD Dissertation, University of Illinois, Urbana, Illinois, 1969.

19. Hanson, J.L., "Linear Difference Games with Delays," PhD Dissertation, University of Illinois, Urbana, Illinois, 1970.

PATTERN RECOGNITION IN REMOTE SENSING[†]

K. S. FU
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

ABSTRACT

This report discusses the application of pattern recognition techniques to the problem of classifying remotely sensed multispectral data. Problems treated include (1) parametric and nonparametric pattern classification and (2) feature extraction and selection. The implementation of these techniques are also discussed. Experimental results from the classifications of agricultural crops are used to demonstrate the applications.

I. INTRODUCTION

Remote sensing technology is concerned with the determination of characteristics of physical objects through the analysis of measurements taken at a distance from these objects. One of the major problems in remote sensing is the characterization and classification of measurements taken from various situations on the Earth's surface. This problem essentially falls into the general problem of pattern recognition [1-5]. The spectral, spatial, and temporal variations of solar energy reflected and absorbed by physical objects are a function of their characteristics. For example, in agricultural situations, remote multispectral sensing is used to detect and record reflected and emitted (electromagnetic radiation) energy from specific target areas in many discrete, relatively narrow spectral bands in several regions of the electromagnetic spectrum [6-7]. Frequently, these lie between 0.3 and 14 microns wavelength. Such bands of electromagnetic radiation may be sensed and recorded using one or a combination of several types of devices, including cameras with diverse film-filter combinations, optical-mechanical scanners with data stored on magnetic tapes, and radar systems. Based on the multispectral data, the problem now is to characterize and classify the data from the agricultural viewpoint.

Most of the early work in the area of remote sensing has been done with image-oriented information. It has been possible to demonstrate adequately the feasibility of a wide variety of identification tasks from the spatial and spectral information which is present in, for example, color photography. The advantage to image-oriented systems is that pictorial information is very well mated to man's capability in terms of data analysis by photointerpretative techniques. Recent developments in the research of image processing by digitalcomputers and/or (analog) optical devices have also increased the possibility of implementing a completely automated image-oriented system [8-12]. On the other hand, numerically oriented information representing the spectral, spatial and temporal variations of electromagnetic fields eminating from the Earth's surface can also be used effectively for the same purpose. From the viewpoint of pattern recognition, this is a case of one problem with two different forms of input data, pictorial and numerical. However, with the present state of the art, the approaches used for these two cases are sometimes quite different. In this paper, only the case of numerically-oriented systems is discussed.

A pattern recognition system can be divided into two parts (see Fig. 1), namely, the feature extractor and the classifier.[††] The function of the

---

[†] This work was supported in part by NSF Grant GK-18225.

[††] This division is primarily for convenience rather than necessity.

feature extractor is to extract or to measure the important characteristics from the input patterns (for example, various scenes on the Earth's surface). The extracted characteristic measurements are called features, and they are supposed to best characterize all the possible input patterns for classification purposes. In other words, the features extracted are presumably relevant for the discrimination of the scenes which we would like to classify. The number of possible scenes (pattern classes) specified in the classification is usually rather subjective. For example, in discriminating green vegetation from non-green vegetation (bare soil, water, etc.) in one area, there are only two pattern classes. However, if we would like to discriminate among five different crops (within green vegetation) and non-green vegetation, then there will be six pattern classes under consideration.

The function performed by a classifier is that of making a proper decision or classification to assign each input pattern to one of the possible pattern classes. The decision is made on the basis of the feature measurements supplied by the feature extractor. The classification criterion is usually the minimization of misclassifications (or misrecognitions). Many classification techniques have recently been proposed. Depending upon whether the feature measurements describing all the possible input patterns in each class can be characterized by deterministic quantities (functions) or statistical quantities (e.g., probability distribution), these techniques can be classified into deterministic and statistical classification techniques. On the other hand, depending upon whether or not the characteristics of the feature measurements describing all the patterns in each class can be expressed in parametric form (for example, by a known form of probability density function), these techniques can be classified into parametric and nonparametric classification techniques. The selection of a particular technique in practical applications is more or less dependent on the nature of the problem, a priori information available, and the designer's preference. Assume that there are m possible pattern classes $\omega_1, \omega_2, \ldots, \omega_m$, and there are N features $x_1, x_2, \ldots, x_N$ to be extracted for classification. Each set of N feature measurements can be represented as an N-dimensional feature vector,

$$X^T = [x_1, x_2, \ldots, x_N]$$

or as a point in the N-dimensional space, called feature space Ω_X. Usually, use of a large number of feature measurements will increase the complexity and computation time of the classifier. In order to have an efficient operation of a recognition system, in addition to the problem of what feature measurements should be taken from the input patterns--feature extraction, one problem raised in practice is to select a subset of smaller size from a given set of available features--feature selection [13-16]. In the following sections, several important techniques in pattern classification and feature selection are briefly discussed, and some of their possible applications in remote sensing problems are illustrated.

II. PARAMETRIC CLASSIFICATION TECHNIQUES

2.1 Bayes Classification Techniques

The concept of pattern classification may be expressed in terms of the partition of feature space or, in terms of statistical decision theory, a testing of statistical hypotheses. A typical two-dimensional three-class illustration of the concept is shown in Fig. 2. The feature space is partitioned into three mutually exclusive regions such that points (patterns) in each region are associated with a particular pattern class. The partitioning boundaries are called decision boundaries or decision surfaces. The determination of the decision boundaries also determines the complete partition of feature space. The problem of classification is to assign each possible vector or point in the feature space (representing a pattern) to a proper pattern class. Since, in many cases, the patterns in one class are expected to have large variations in feature measurements and the noise effect involved in taking these measurements cannot be neglected, one approach proposed is to consider that $x_1, x_2, \ldots, x_N$, are random variables where x_i is the noisy measurement of the ith feature. For each pattern class ω_j, $j=1,\ldots,m$, assume that the multivariate (N-dimensional) probability density (or distribution) function of the feature vector X, $p(X/\omega_j)$, and the probability of occurrences of ω_j, $P(\omega_j)$, are known or can be estimated. On the basis of the a priori information $p(X/\omega_j)$ and $P(\omega_j)$, $j=1,\ldots,m$, the function of a classifier is to perform the classification task for minimizing probability of misrecognition. The problem of pattern classification can now be formulated as a statistical decision problem (testing of m statistical hypotheses) by defining a decision function $\delta(X)$ where $\delta(X) = \delta_i$ means that the hypothesis H_i: $X \sim \omega_i$ (X is from the class ω_i) is accepted. Let $L(\omega_i, \delta_j)$ be the loss incurred by the classifier if the decision δ_j is made when the input pattern is actually from ω_i. The conditional loss (or conditional risk) is

$$r(\omega_i, \delta) = \int_{\Omega_X} L(\omega_i, \delta)\, p(X/\omega_i)\, dX \tag{2.1}$$

For a given set of a priori probabilities $P = \{P(\omega_1), P(\omega_2), \ldots, P(\omega_m)\}$, the average or expected loss (or average risk) is

$$R(P, \delta) = \sum_{i=1}^{m} P(\omega_i)\, r(\omega_i, \delta) \tag{2.2}$$

Substitute (2.1) into (2.2) and let

$$r_X(P, \delta) = \frac{\sum_{i=1}^{m} L(\omega_i, \delta)\, P(\omega_i)\, p(X/\omega_i)}{p(X)} \tag{2.3}$$

Then (2.2) becomes

$$R(P, \delta) = \int_{\Omega_X} p(X)\, r_X(P, \delta)\, dX \tag{2.4}$$

$r_X(P, \delta)$ is defined as the a posteriori conditional average loss of the decision δ for given feature measurements X.

The problem is to choose a proper decision δ_j, $j=1,\ldots,m$, to minimize the average loss or risk $R(P, \delta)$, or to minimize the maximum of the conditional average loss $r(\omega_i, \delta)$ (minimax criterion[†]). The optimal decision rule which minimizes the average loss is called the Bayes' rule. From (2.4) it is sufficient to consider each X separately and to minimize $r_X(P, \delta)$. If δ^* is an optimal decision in the sense of minimizing the average loss, then

$$r_X(P, \delta^*) \leq r_X(P, \delta) \tag{2.5}$$

that is,

$$\sum_{i=1}^{m} L(\omega_i, \delta^*)\, p(X/\omega_i) \leq \sum_{i=1}^{m} L(\omega_i, \delta)\, P(\omega_i)\, p(X/\omega_i) \tag{2.6}$$

For the (0,1) loss function, i.e.,

[†] In some classification problems, the information about a priori probabilities is not available. The minimax criterion (with respect to the least favorable a priori distribution) is suggested as a classification procedure.

$$L(\omega_i, \delta_j) = \begin{cases} 0, & i=j \\ 1, & i\neq j \end{cases} \qquad (2.7)$$

the average loss is essentially also the probability of misclassification. In this case, the Bayes' decision rule is that $\delta^* = \delta_i$, i.e., $X \sim \omega_i$ if

$$P(\omega_i)p(X/\omega_i) \geq P(\omega_j)p(Z/\omega_j) \text{ for all } j=1,\ldots,m. \qquad (2.8)$$

Define the likelihood ratio between class ω_i and class ω_j as

$$\lambda = p(X/\omega_i)/p(X/\omega_j) \qquad (2.9)$$

Then (2.8) becomes

$$\delta^* = \delta_i \quad \text{if } \lambda \geq P(\omega_j)/P(\omega_i) \quad \text{for all } j=1,\ldots,m. \qquad (2.10)$$

A classifier implementing the Bayes' decision rule for classification is called a Bayes classifier. A simplified block diagram of a Bayes classifier is shown in Fig. 3. The decision boundary between regions in Ω_X associated with ω_i and ω_j is

$$P(\omega_i)\, p(X/\omega_i) - P(\omega_j)\, p(X/\omega_j) = 0 \qquad (2.11)$$

or

$$\log \frac{P(\omega_i)\, p(X/\omega_i)}{P(\omega_j)\, p(X/\omega_j)} = 0 \qquad (2.12)$$

As an illustrative example, suppose that $p(X/\omega_i)$, $i=1,\ldots,m$, is a multivariate gaussian density function with mean vector M_i and covariance matrix K_i, i.e.,

$$p(X/\omega_i) = \frac{1}{(2\pi)^{N/2}|K_i|^{1/2}} \exp\left[-\frac{1}{2}(X-M_i)^T K_i^{-1}(X-M_i)\right],^{\dagger} \quad i=1,\ldots,m \qquad (2.13)$$

Then the decision boundary expressed by (2.12) is

$$\log \frac{P(\omega_i)}{P(\omega_j)} + \log \frac{p(X/\omega_i)}{p(X/\omega_j)} = \log \frac{P(\omega_i)}{P(\omega_j)} - \frac{1}{2}\left[(X-M_i)^T K_i^{-1}(X-M_i) - (X-M_j)^T K_j^{-1}(X-M_j)\right]$$
$$= 0 \qquad (2.14)$$

Equation (2.14) is, in general, a hyperquadrics (a quadric function in high dimensional space). If $K_i=K_j=K$, (2.14) reduces to

$$X^T K^{-1}(M_i-M_j) - \frac{1}{2}(M_i+M_j)^T K^{-1}(M_i-M_j) + \log \frac{P(\omega_i)}{P(\omega_j)} = 0 \qquad (2.15)$$

which is a hyperplane.

It is noted that from (2.8), $P(\omega_i)p(X/\omega_i) = p(\omega_i,X) = P(\omega_i/X)p(X)$. Hence, the Bayes' decision rule with (0,1) loss function is also the unconditional maximum-likelihood decision rule (MLDR). Furthermore, the (conditional) maximum-likelihood decision may be regarded as the Bayes' decision rule (2.8) with equal a priori probabilities, i.e., $P(\omega_i) = 1/m$, $i=1,\ldots,m$.

In the Bayes classification system described above, all the N features are observed by the classifier at one stage. As a matter of fact, the cost of feature measurements has not been taken into consideration. It is evident that an insufficient number of feature measurements will not be able to give satisfactory results in correct classification. On the other hand, an arbitrarily large number of features to be measured is impractical. If the cost of taking feature measurements is to be considered or if the features extracted from input patterns are sequential in nature, one is led to apply sequential decision procedures to this class of pattern recognition problems. The problem is especially pertinent when the cost of taking a feature

---

[†] The superscript T represents the transpose operation to a vector.

measurement is high. A tradeoff between the error (misrecognition) and the number of features to be measured can be obtained by taking feature measurements sequentially and terminating the sequential process (making a decision) as soon as a sufficient or desirable accuracy of classification has been achieved.† Examples of the sequential classification procedure applied to agricultural remote sensing problems can be found in [17].

2.2 Classification Based on the Distances Between Class Distributions

In some remote sensing problems, classification can be carried out on the per-field basis. For example, in accordance with general crop planting practice, if an agricultural field has a certain majority of resolution elements classified as "ω_i", then one can feel reasonably sure that the entire field as a whole is in class "ω_i". Based on this observation, we shall formulate a pattern classification procedure as follows [18]. Let $F_1, F_2, \ldots, F_m$ be the probability distributions of the feature vectors corresponding to the pattern classes $\omega_1, \omega_2, \ldots, \omega_m$, respectively. Then, given a set of points (patterns) $\{X_1, X_2, \ldots, X_n\} = \{X\}$ all from the same field, the problem is to decide to which F_i they belong. A distance measure $d(F_i, F_j)$ is proposed as a suitable metric of the separation between any two distributions F_i and F_j, and this metric will be used as a rule for decision making. Let $G(X)$ be the distribution governing $\{X\}$. We shall compare the magnitudes of $\{d(G, F_i)\}$; the distribution F_i which minimizes the distance $d(G, F_i)$ shall be presumed to be the distribution which contains $G(X)$. Therefore, the vector set $\{X\}$ is classified into one of the known classes.

Let $F_i(X)$ and $F_j(X)$ be the distributions of the pattern classes ω_i and ω_j, respectively. Then the distance between distributions F_i and F_j is defined as follows:

$$d^2(F_i, F_j) = \int_{\Omega_X} \left(\sqrt{p(X/\omega_i)} - \sqrt{p(X/\omega_j)} \right)^2 dX \tag{2.16}$$

If we define

$$\rho(F_i, F_j) = \int_{\Omega_X} \sqrt{p(X/\omega_i)} \cdot \sqrt{p(X/\omega_j)}\, dX \,, \tag{2.17}$$

we have

$$d^2(F_i, F_j) = 2 - 2\rho(F_i, F_j) \tag{2.18}$$

The quantity $\rho(F_i, F_j)$ expresses the correlation between distributions F_i and F_j, and we can use $\rho(F_i, F_j)$ in decision making. It is important to note that to minimize the $d(F_i, F_j)$ is the same as to maximize the $\rho(F_i, F_j)$. Hence, the decision (classification) can be made by choosing the class which maximizes the correlation function ρ.

The distance $d(F_i, F_j)$ defined in (2.16) is applicable to any pair of distributions. We shall now turn to a specific example of multivariate gaussian case. Let F_i and F_j be N-dimensional gaussian distributions with their probability density functions given in (2.13). From (2.17), we obtain

$$\rho(F_i, F_j) = \frac{|K_i^{-1} K_j^{-1}|^{1/4}}{|\frac{1}{2}(K_i^{-1} + K_j^{-1})|^{1/2}} \exp\left[\frac{1}{4} (K_i^{-1} M_i + K_j^{-1} M_j)^T \cdot \right.$$

$$\left. (K_i^{-1} + K_j^{-1})^{-1} (K_i^{-1} M_i + K_j^{-1} M_j) - \frac{1}{4} (M_i^T K_i^{-1} M_i + M_j^T K_j^{-1} M_i) \right] \tag{2.19}$$

When $K_i = K_j = K$,

---

† This condition is often required in target and disease detection problems.

$$\rho(F_i,F_j) = \exp\left[-\frac{1}{8}(M_i-M_j)^T K^{-1}(M_i-M_j)\right]$$

and when $M_i = M_j = M$,

$$\rho(F_i,F_j) = \frac{|K_i^{-1}\ K_j^{-1}|^{1/4}}{|\frac{1}{2}(K_i^{-1}+K_j^{-1})|^{1/2}} \tag{2.20}$$

2.3 Illustrative Examples

The data used in the following examples were measured and recorded from aircraft flights at 1500 - 7500 feet.† Optical-mechanical scanners are used to measure the energy reflectance from the ground in twelve different wavelength bands. These wavelength bands include the following: 0.40-0.44μ, 0.44-0.46μ, 0.46-0.48μ, 0.48-0.50μ, 0.50-0.52μ, 0.52-0.55μ, 0.55-0.58μ, 0.58-0.62μ, 0.62-0.66μ, 0.66-0.72μ, 0.72-0.80μ, and 0.80-1.0μ. The last two wavelength bands (channels 11 and 12) are in the reflective infrared portion of the spectrum, and the other bands encompass the visible wavelengths. These twelve measurements are recorded on a twelve-channel magnetic tape, and they constitute the basic twelve feature measurements for classification purposes. The twelve-channel recorded data are then digitized for the purpose of digital processing.

For the purpose of crop classification, the statistical properties of the twelve-dimensional feature measurement vectors or samples for each pattern class (each kind of crops or ground cover, in this case) were investigated. Ground truth information was provided initially for the (training) samples which were used to estimate the statistical characteristics for each class. Univariate histograms were established for each class. It was noted that, in most cases, the histograms were unimodal in shape. It seemed reasonable, during the initial phase of study, to approximate the conditional probability density functions of the feature measurements for each class by multivariate gaussian density functions. The mean vectors and covariance matrices for each class were estimated by means of the sample mean vector and the sample covariance matrix calculated from training samples. In the special case of wheat, the histograms established indicated bimodal shapes. The bimodal density function was hypothesized as a mixture of two gaussian density functions, each of which represented a subclass of wheat crops. Therefore, the assumption of gaussian distributed feature measurements is still a reasonably good approximation if the subclasses can be appropriately selected. Typical histograms for wheat are shown in Fig. 4. For m pattern classes (e.g., m kinds of agricultural crops) $\omega_1,\omega_2,\ldots,\omega_m$, from the gaussian assumption, the feature measurement vectors for each class are distributed according to (2.13) where X is a 12-dimensional vector and M_i and K_i are now the sample mean vector and the sample covariance matrix for the ith class, ω_i.

Example 1: Maximum Likelihood Classification. The maximum likelihood classification rule (or Bayes decision rule with (0,1) loss function) is used for the classification task. Following (2.12), let

$$g_i(X) = \log P(\omega_i)p(X/\omega_i),\ i=1,\ldots,m \tag{2.21}$$

Assume equal a priori probabilities, with gaussian assumption

$$g_i(X) = \log P(\omega_i) - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|K_i| - \frac{1}{2}\left[(X-M_i)^T K_i^{-1}(X-M_i)\right] \qquad i-1,\ldots,m \tag{2.22}$$

or simply,

---

† The details concerning how to take the measurements are beyond the scope of this paper.

$$g_i(X) = -\frac{1}{2}\log|K_i| - \frac{1}{2}[(X-M_i)^T K_i^{-1}(X-M_i)],\ i=1,\ldots,m \quad (2.23)$$

The classification rule is then reduced to: classify X as from the class ω_i if

$$\log|K_i| + (X-M_i)^T K_i^{-1}(X-M_i) < \log|K_j| + (X-M_j)^T K_j^{-1}(X-M_j),\ \text{for all } j\neq i \quad (2.24)$$

The classification scheme was tested for a nine-class crop classification problem. The data were taken from an area in west-central Tippecanoe County, Indiana (Flight Line C1). The area is approximately one mile by four miles in size. The nine classes are rye, oats, corn, soybeans, alfalfa, red clover, bare soil and two classes of wheat. The classification results using all twelve features (bands) are given in Tables 1 and 2. In order to test the classifier's capability of generalization, the training samples were also used as test samples for classification in Table 1, but a completely different set of test samples for classification were used in Table 2. The difference in classification accuracy is probably due to the fact that the training samples used were not completely representative and the number of training samples was inadequate.

Example 2: Per-Field Classification Based on the Distances Between Distributions. In order to evaluate the classification scheme discussed in Section 2.2, 81 agriculture fields in the Flight Line C4 area were used as test fields. Based on the histogram study, five subclasses were chosen to represent the soybean, the corn and the pasture. Four subclasses were used to represent stubble. This was due to the multimodal distribution of each of the four classes. The classification results are summarized in Table 3. When the same data was classified by the Maximum-Likelihood Classification Rule (MLDR) with the same set of training statistics, the classification accuracy was 79.7 percent.

III. NONPARAMETRIC CLASSIFICATION TECHNIQUES

In the classification techniques discussed in the previous sections, the knowledge about $p(X/\omega_i)$, $i=1,\ldots,m$, is assumed known a priori or can be reasonably accurately estimated by learning. When an a priori assumption (for example, gaussian assumption) characterizing the patterns under study cannot be made with some degree of validity, or the estimation (or learning) of unknown information cannot be made with some degree of accuracy (say, due to insufficient learning samples), it appears questionable to blindly apply parametric techniques. Instead, nonparametric techniques must be used. In the following sections, several nonparametric classification techniques are briefly introduced and some of their applications to crop classification problems illustrated.

3.1 Discriminant Functions Approach

Based on the concept of partition of the feature space, the problem of classification can also be mathematically formulated in terms of "discriminant functions" [2]. The discriminant function $D_j(X)$ associated with pattern class ω_j, $j=1,\ldots,m$, is defined such that if $X \sim \omega_i$, the value of $D_i(X)$ must be the largest. That is, for all $X \sim \omega_i$,

$$D_i(X) > D_j(X),\ i,j=1,\ldots,m,\quad i\neq j \quad (3.1)$$

Thus, in the feature space Ω_X, the decision boundary between the regions associated with class ω_i and class ω_j is expressed by the following equation:

$$D_i(X) - D_j(X) = 0 \quad (3.2)$$

A general block diagram for the classifier using criterion (3.1) is shown in Fig. 5. Many different forms satisfying condition (3.1) can be selected for $D_i(X)$. Several important discriminant functions are discussed in the following.

(A) Linear Discriminant Functions. In this case, a linear combination of the feature measurements $x_1, x_2, \ldots, x_N$ is selected for $D_i(X)$; i.e.,

$$D_i(X) = \sum_{k=1}^{N} w_{ik}x_k + w_{i,N+1}, \quad i=1,\ldots,m. \tag{3.3}$$

The decision boundary between regions in Ω_X associated with ω_i and ω_j is in the form of

$$D_i(X) - D_j(X) = \sum_{k=1}^{N} w_k x_k + w_{N+1} = 0 \tag{3.4}$$

with $w_k = w_{ik} - w_{jk}$ and $w_{N+1} = w_{i,N+1} - w_{j,N+1}$. Equation (3.4) is the equation of a hyperplane in the feature space Ω_X. If $m=2$, on the basis of (3.4), $i,j = 1,2$ $(i \neq j)$, a threshold logic device as shown in Fig. 6 can be employed as a linear classifier (a classifier using linear discriminant functions). From Fig. 6, let $D(X) = D_1(X) - D_2(X)$, if

$$\begin{aligned} &\text{output} = +1, \text{ i.e., } D(X) > 0,\ X \sim \omega_1; \text{ and if} \\ &\text{output} = -1, \text{ i.e., } D(X) < 0,\ X \sim \omega_2 \end{aligned} \tag{3.5}$$

For the number of pattern classes more than two $(m > 2)$, several threshold logic devices can be connected in parallel so that the combinations of the outputs from, say, M threshold logic devices will be sufficient for distinguishing m classes, i.e., $2^M \geq m$. Or the general configuration of Fig. 5 can also be used.

(B) Minimum-Distance Classifier. An important class of linear classifiers is that of using the distances between the input pattern and a set of reference vectors or prototype points in the feature space as the classification criterion. Suppose that m reference vectors $R_1, R_2, \ldots, R_m$, are given with R_j associated with the pattern class ω_j. A minimum-distance classification scheme with respect to $R_1, R_2, \ldots, R_m$, is to classify the input X as from class ω_i; i.e.,

$$X \sim \omega_i \quad \text{if} \quad d(X,R_i) \text{ is the minimum,} \tag{3.6}$$

where $d(X,R_i)$ is the distance defined between X and R_i. For example, $d(X,R_i)$ may be defined as

$$d(X,R_i) = \sqrt{(X-R_i)^T(X-R_i)} \tag{3.7}$$

From (3.7),

$$d^2(X,R_i) = X^TX - X^TR_i - XR_i^T + R_i^TR_i \tag{3.8}$$

Since X^TX is not a function of i, the corresponding discriminant function for a minimum-distance classifier is essentially

$$D_i(X) = X^TR_i + XR_i^T - R_i^TR_i, \quad i=1,\ldots,m \tag{3.9}$$

which is linear. Hence, a minimum-distance classifier is also a linear classifier. The performance of a minimum-distance classifier is, of course, dependent upon an appropriately selected set of reference vectors.

(C) Piecewise Linear Discriminant Functions. The concept adopted in (B) can be extended to the case of minimum-distance classification with respect to sets of reference vectors. Let $R_1, R_2, \ldots, R_m$ be the m sets of reference vectors associated with classes $\omega_1, \omega_2, \ldots, \omega_m$, respectively, and let reference vectors in R_j be denoted as $R_j^{(k)}$; i.e.,

$$R_j^{(k)} \in R_j, \quad k=1,\ldots,u_j,$$

where u_j is the number of reference vectors in set R_j. Define the distance between an input feature vector X and R_j as

$$d(X,R_j) = \min_{k=1,\ldots,u_j} \{d(X,R_j^{(k)})\} \tag{3.10}$$

That is, the distance between X and R_j is the smallest of the distances between X and each vector in R_j. The classifier will assign the input to a pattern class which is associated with the closest vector set. If the distance between X and $R_j^{(k)}$, $d(X,R_j^{(k)})$ is defined as (3.7), then the discriminant function used in this case is essentially

$$D_i(X) = \max_{k=1,\ldots,u_i} \left\{ X^T R_i^{(k)} + (R_i^{(k)})^T X - (R_i^{(k)})^T R_i^{(k)} \right\}, \quad i=1,\ldots,m \tag{3.11}$$

Let

$$D_i^{(k)} = X^T R_i^{(k)} + (R_i^{(k)})^T X - (R_i^{(k)})^T R_i^{(k)} \tag{3.12}$$

Then

$$D_i(X) = \max_{k=1,\ldots,u_i} \{D_i^{(k)}(X)\}, \quad i=1,\ldots,m \tag{3.13}$$

It is noted that $D_i^{(k)}(X)$ is a linear combination of features; hence, the class of classifiers using (3.11) or (3.13) is often called piecewise linear classifiers. An example of the piecewise linear classifier is the α-perceptron [19].

(D) Polynomial Discriminant Functions [20,21]. An rth-order polynomial discriminant function can be expressed as

$$D_i(X) = w_{i1} f_1(X) + w_{i2} f_2(X) + \ldots + w_{iL} f_L(X) + w_{1,L+1} \tag{3.14}$$

where $f_j(X)$ is of the form

$$x_{k_1}^{n_1} x_{k_2}^{n_2} \ldots x_{k_r}^{n_r} \text{ for } k_1, k_2, \ldots, k_r = 1,\ldots,N, \text{ and } n_1, n_2, \ldots, n_r = 0 \text{ and } 1 \tag{3.15}$$

The decision boundary between any two classes is also in the form of an rth-order polynomial. Particularly, if r=2, the discriminant function is called a quadric discriminant function. In this case,

$$f_j(X) = x_{k_1}^{n_1} x_{k_2}^{n_2} \text{ for } k_1, k_2 = 1,\ldots,N, \text{ and } n_1, n_2 = 0 \text{ and } 1 \tag{3.16}$$

and

$$L = \frac{1}{2} N(N+3) \tag{3.17}$$

Typically,

$$D_i(X) = \sum_{k=1}^{N} w_{kk} x_k^2 + \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} w_{jk} x_j x_k + \sum_{j=1}^{N} w_j x_j + w_{L+1} \tag{3.18}$$

In general, the decision boundary for quadric discriminant functions is a hyperhyperboloid. Special cases include hypersphere, hyperellipsoid and hyperellipsoidal cylinder. A general quadric discriminant computer is shown in Fig. 7.

It is noted from (2.8) that the corresponding discriminant function implemented by a Bayes classifier is essentially

$$D_i(X) = P(\omega_i)\, p(X/\omega_i), \quad i=1,\ldots,m \tag{3.19}$$

or equivalently,

$$D_i(X) = \log P(\omega_i)\, p(X/\omega_i), \quad i=1,\ldots,m \tag{3.20}$$

This gives an alternative interpretation of the decision rule (2.8) in terms of discriminant functions. Also, the $g_i(X)$ selected in (2.21) is precisely the same as the discriminant function $D_i(X)$ defined in (3.20).

3.2 Training in Linear Classifiers

The two-class linear classifier discussed in Section 3.1(A) is easily implemented by a single threshold logic device. If the patterns from different classes are linearly separable (can be separated by a hyperplane in the feature space Ω_X), then, with correct values of the coefficients or weights, $w_1, w_2, \ldots, w_{N+1}$, in (3.4), a perfectly correct recognition is possible to be achieved. However, in practice, the proper values of the weights are usually not available. Under such circumstances, it is proposed that the classifier is designed to have the capability of estimating the best values of the weights from the input patterns. The basic idea is that, by observing patterns with known classifications, the classifier can automatically adjust the weights in order to achieve correct recognitions. The performance of the classifier is supposed to improve as more and more patterns are applied. This process is called training or learning, and the patterns used as the inputs are called training patterns. Several simple training rules are briefly introduced in this section.

Let Y be an augmented feature vector which is defined as

$$Y^T = [x_1, x_2, \ldots, x_N, 1] = [X^T, 1] \tag{3.21}$$

where X is the feature vector of a pattern. Consider two sets of training patterns T_1 and T_2 belonging to two different pattern classes ω_1 and ω_2, respectively. Corresponding to the two training sets, there are two sets of augmented vectors T_1' and T_2'. Each element in T_1' and T_2' is obtained by augmenting the patterns in T_1 and T_2, respectively. The two training sets are linearly separable which means that a weight vector W exists (called the solution weight vector) such that

$$Y^T W > 0 \text{ for each } Y \in T_1' \quad \text{or} \quad Y^T W < 0 \text{ for each } Y \in T_2' \tag{3.22}$$

where

$$W^T = [w_1, w_2, \ldots, w_{N+1}] \tag{3.23}$$

The so-called "error-correction" training procedure for a linear classifier can be summarized as follows. For any $Y \in T_1'$, the product $Y^T W$ must be positive; i.e., $Y^T W > 0$. If the output of the classifier is erroneous (i.e., $Y^T W < 0$) or undefined (i.e., $Y^T W = 0$), then let the new weight vector be

$$W' = W + \alpha Y \tag{3.24}$$

where $\alpha > 0$ is called the correction increment. On the other hand, for $Y \in T_2'$, $Y^T W < 0$. If the output of the device is erroneous (i.e., $Y^T W > 0$) or undefined, then let

$$W' = W - \alpha Y \tag{3.25}$$

Before training begins, W may be preset to any convenient values. Three rules for choosing α are suggested: (i) Fixed increment rule--α is any fixed positive number; (ii) absolute correction rule--α is taken to be the smallest integer which will make the value of $Y^T W$ cross the threshold of zero. That is,

$$\alpha = \text{the smallest integer greater than } |Y^T W| / Y^T Y \; ; \tag{3.26}$$

(iii) fractional correction rule--α is chosen such that

$$|Y^T W - Y^T W'| = \lambda |Y^T W|, \quad 0 < \lambda \leq 2; \tag{3.27}$$

or, equivalently,

$$\alpha = \lambda \;\; |Y^T W| / Y^T Y \tag{3.28}$$

The convergence of the three error-correction rules can be proved [2]. By convergence, we mean that if the two training sets are linearly separable, the sequence of weight vectors produced by the training rule converges to a solution weight vector in a finite number of training steps or iterations.

3.3 Potential Function Method

One nonparametric approach which has been investigated for its applicability to crop classification is the Potential Function Method [22]. The method is briefly reviewed in the following. Let $\varphi_i(X)$, i=1,2,..., be a complete set of functions defined on the feature space (e.g., a set of orthonormal functions). Suppose that an unknown function f(X) to be estimated from training samples can be represented by the expansion (3.29). The unknown function, in this application, may be a discriminant function or a probability density function. A function of two variables,

$$f(X) = \sum_{i=1}^{M} c_i \varphi_i(X) \tag{3.29}$$

called the potential function, is introduced as

$$K(X,Y) = \sum_{i=1}^{M} \lambda_i^2 \varphi_i(X) \varphi_i(Y) \tag{3.30}$$

where λ_i's are real numbers chosen in such a way that the function K(X,Y) is bounded. After the observation of n training samples, $X_1, X_2, \ldots, X_n$, the nth estimate of the function f(X), denoted by $f_n(X)$ can be computed from the following general algorithm:

$$f_n(X) = f_{n-1}(X) + r_n K(X, X_n) \tag{3.31}$$

with $f_o(X) \equiv 0$. r_n is dependent upon the type of information received about $f(X_n)$ and the area of application.

In the estimation of probability density function of the feature measurements for each pattern class (each class of crop), let the density function p(X) be represented by

$$p(X) = \sum_{i=1}^{M} c_i \varphi_i(X) \tag{3.32}$$

Let

$$K(X,Y) = \sum_{i=1}^{M} \varphi_i(X) \varphi_i(Y) \tag{3.33}$$

where $\varphi_i(X)$, i=1,...,M, are bounded. After the training samples $X_1, \ldots, X_n$ are taken, the nth estimate of the probability density function can be directly expressed as

$$p_n(X) = \frac{1}{n} \sum_{k=1}^{n} K(X, X_k) \tag{3.34}$$

It has been proved that, using (3.34), $p_n(X)$ converges to p(X) in the mean square sense. An alternative form of (3.34) can be obtained by substituting (3.33) into (3.34)

$$p_n(X) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{M} \varphi_i(X) \varphi_i(X_k). \tag{3.35}$$

Defining

$$c_{i,n} = \frac{1}{n} \sum_{k=1}^{n} \varphi_i(X_k) \tag{3.36}$$

yields the nth estimate of p(X).

$$p_n(X) = \sum_{i=1}^{M} c_{i,n} \varphi_i(X) \tag{3.37}$$

where

$$c_{i,n} = c_{i,n-1} + \frac{1}{n} [\varphi_i(X_n) - c_{i,n-1}], \quad n=1,2,\ldots \tag{3.38}$$

and $c_{i,o} = 0$.

In the application of the Potential Function Method to multiclass classification problems, let the unknown function f(X) be the discriminant function for each class. After the training samples $X_1, X_2, \ldots, X_n$ are taken, let X_n be coming from the class ω_i. The algorithm is that

(a) if $D_{i,n-1}(X_n) > D_{j,n-1}(X_n)$ for some $j \neq i$ (3.39)

then $D_{i,n}(X) = D_{i,n-1}(X)$; (3.40)

(b) if $D_{k,n-1}(X_n) \geq D_{i,n-1}(X_n)$ for some $k \neq i$ (3.41)

then $D_{k,n}(X) = D_{k,n-1}(X) - \alpha\, K(X,X_n)$,

$D_{i,n}(X) = D_{i,n-1}(X) + \alpha\, K(X,X_n)$, and

$D_{\ell,n}(X) = D_{\ell,n-1}(X)$ $\quad \ell \neq i,\ \ell \neq k$ (3.42)

where α is an adjustable parameter. It is obvious that (3.42) is a special case of (3.31). In a special case of two-class classification problems, let

$$f_n(X) = D_{1,n}(X) - D_{2,n}(X) \tag{3.43}$$

If a training sample X_n from class ω_1 is misclassified as from class ω_2, then

$$D_{1,n}(X) = D_{1,n-1}(X) + \alpha\, K(X,X_n) \tag{3.44}$$

and

$$D_{2,n}(X) = D_{2,n-1}(X) - \alpha\, K(X,X_n) \tag{3.45}$$

From (3.43), (3.44) and (3.45),

$$f_n(X) = f_{n-1}(X) + 2\,\alpha\, K(X,X_n) \tag{3.46}$$

The Potential Function Method just described was applied to a two-class crop classification problem [23]. The two classes of crops were corn and soybeans. The test data were taken from Flight Line C2. Features 1 (0.40-0.44μ), 10 (0.66-0.72μ), and 12 (0.80-1.00μ) were used. A total of 912 samples (456 from each class) were tested. The histogram analysis indicated the existence of multimodal shapes. This means that the gaussian assumption used previously is probably not a good approximation. Two hundred samples from each class were used as training samples. For the purpose of comparison, three different classification schemes were tested.

(1) The gaussian assumption for each class is used. The mean vectors and covariance matrices were estimated from the training samples. The 912 samples were classified by the maximum likelihood classification rule. The a priori probabilities are assumed equal.

(2) The probability density function of feature measurements for each class was estimated from the training samples using the Potential Function Method. The potential function used is of the following form:

$$K(X,Y) = e^{(X-Y)^T(X-Y)} \tag{3.47}$$

The 912 samples were classified by the maximum likelihood classification rule based on the estimated density functions.

(3) Nonlinear discriminant functions were used to form the decision function defined by (3.43). The Potential Function Method with $K(X,Y)$ given in (3.47) was applied to estimate the decision function $f(X) = D_1(X) - D_2(X)$. Training samples from the two classes were alternatively presented. Four cycles through the training samples were made during training. Ten misclassifications still remained. These were due to identical sample vectors being in different classes. The final form of $f_n(X)$ consisted of 101 terms. Classification of the 912 samples was performed on the basis of the sign of $f_n(X)$.

The classification results for the 912 samples are:

Classification scheme 1:	76.4% (215 misclassifications)
Classification scheme 2:	91.2% (81 misclassifications)
Classification scheme 3(α=1):	88.4% (108 misclassifications)

The classification results for training samples only are:

Classification scheme 1: 81.7% (73 misclassifications)
Classification scheme 2: 98.3% (7 misclassifications)
Classification scheme 3: 97.5% (10 misclassifications)

It is noted that, from the results shown above, the classification accuracy is higher in the case where the Potential Function Method is used. In general, compared with Classification scheme 3, Classification scheme 2 requires more storage and computation time.

3.4 Nearest-Neighbor Classification Rule

Let $T = \{X_1, \ldots, X_n\}$ be a set of n training samples, and let $X_n' \in T$ be a sample (pattern) nearest to X. Then the "nearest-neighbor" (NN) classification rule for classifying X is to assign it to the pattern class associated with X_n' [24]. That is, X_n' is a nearest neighbor to X if

$$\text{Min}\ d(X_k, X) = d(X_n', X)\ , \qquad k=1,\ldots,n \tag{3.48}$$

and by the NN rule, X is assigned to the class of its nearest neighbor X_n'. A misclassification is made if X and X_n' actually belong to two different classes. In the following, the performance of NN classification rule is briefly analyzed.

Let $X \sim \omega$ and $X_n' \sim \omega'$ where $\omega, \omega' \in \{\omega_1, \ldots, \omega_m\}$. According to NN-rule, $\omega = \omega'$ if X_n' is the nearest neighbor to X. Let the loss incurred by the NN-rule be $L(\omega, \omega')$. The n-sample NN risk R(n) is defined by the expectation

$$R(n) = E[L(\omega, \omega')] \tag{3.49}$$

and the (large sample) NN risk R by

$$R = \lim_{n \to \infty} R(n) \tag{3.50}$$

It can be shown that, with (0,1) loss and under certain conditions imposed on the sample statistics [24], $X_n' \to X$ with probability one and

$$R^* \leq R \leq R^*\left(2 - \frac{m}{m-1}\ R^*\right) \tag{3.51}$$

where R^* is the Bayes risk (minimum expected risk or probability of misclassification). Since the Bayes risk is the best performance that can be achieved, the NN-rule is certainly a suboptimal classification procedure. However, with an unlimited number of training samples the NN risk is never worse than twice the Bayes risk. It is easy to see that this upper bound in (3.51) is achieved when the probability density functions $p(X/\omega_i)$, $i=1,\ldots,m$, are identical. Thus the bounds given by (3.51) are as tight as possible, in the sense that for any R^* there exists $p(X/\omega_i)$ and $P(\omega_i)$, $i=1,\ldots,m$, for which the bounds are achieved. Since R is always less than or equal to $2R^*$, even with infinite collection of data (training samples) and as arbitrarily complicated classification rule, one could, at most, cut the risk in half. In this sense, half of the classification information in an infinite set of training samples resides in the nearest neighbor. The nature of (3.51) can be illustrated graphically as shown in Fig. 8. The Bayes risk can be anywhere between 0 and $(m-1)/m$. The bounds meet at these two extreme points. When the Bayes risk is small, the upper bound is approximately twice the Bayes risk. In general, the NN risk must fall in the shaded area shown.

The performance of the NN rule in the finite sample case, like other nonparametric techniques, is rather difficult to evaluate without making further assumptions about the underlying probability structure. The risk or probability of misclassification does not necessarily decrease monotonically with n, and the convergence can be arbitrarily slow in the finite sample case. An obvious extension of the NN classification rule is the k-nearest-neighbor classification rule. The k-NN classification rule classifies X by assigning it the pattern class most frequently represented among the k nearest samples;

in other words, a decision is made by examining the classes of the k nearest neighbors and taking a vote. In order to avoid a tie in the vote taking, k is usually chosen as an odd number and a small fraction of n. The mathematical analysis of k-NN rule is omitted here.

IV. FEATURE EXTRACTION AND SELECTION

As mentioned in Section I, the extraction and selection of features are important problems in pattern recognition and are closely related to the performance of classification. Also, in sequential pattern recognition systems, the ordering of features for successive measurements is important. The purpose of feature ordering is to provide, at successive stages of the sequential classification process, a feature which is the most "informative" among all possible choices of features for the next measurement so the process can be terminated earlier. The problem of feature ordering may be considered to be a problem of feature selection if, at each stage of the sequential classification process, the feature subset to be selected contains only one feature. In the following sections, two major techniques for feature extraction and selection are discussed and their applications to agricultural remote sensing demonstrated.

4.1 Feature Space Transformation

One approach to feature extraction and selection is to extract or select the feature which will "best" characterize the input patterns. The idea is to formulate the problem as a representation or approximation problem. A given feature space is transformed into a new feature space with lower dimensions such that a good representation or approximation of the pattern characteristics can be achieved in the new space. Although, in principle, both linear and nonlinear transformations can be applied, most results obtained so far are primarily concerned with linear transformations.

Let $p(X/\omega)$ be a multivariate probability density function with mean vector M and let Z be the normalized pattern feature vector from class ω; that is,

$$Z^T = X^T - M^T = [z_1, z_2, \ldots, z_N] \tag{4.1}$$

Assume that Z can be expressed by

$$Z = \sum_{k=1}^{N} f_k(X)\, \varphi_k \tag{4.2}$$

where $\{\varphi_k\}$ is a set of orthonormal coordinate vectors with

$$\varphi_k^T = [\varphi_{k1}, \varphi_{k2}, \ldots, \varphi_{kN}], \qquad \varphi_k^T \varphi_i = \begin{cases} 0 \text{ for } k \neq i \\ 1 \text{ for } k = i \end{cases} \tag{4.3}$$

$f_k(X)$, k=1,...,N, are the coefficients in the orthonormal expansion and are the function of X. If we would like to represent or to approximate Z in (4.2) by an ℓ-term summation ($\ell < N$) in the sense of minimizing the mean square error of approximation, then the problem reduces to that of selecting φ_k, k=1,..., ℓ to minimize†

$$R_\ell(\varphi) = E\left[|Z - \sum_{k=1}^{\ell} f_k(X)\varphi_k|^2 \right] = E\left[\sum_{k=\ell+1}^{N} f_k(X)\varphi_k|^2 \right] \tag{4.4}$$

Multiply both sides of (4.2) by φ_i^T and obtain

$$\varphi_i^T Z = \sum_{k=1}^{N} f_k(X)\, \varphi_i^T \varphi_k \tag{4.5}$$

Using the orthonormal condition in (4.3), (4.5) reduces to

$$\varphi_i^T Z = f_i(X) \tag{4.6}$$

---

†The problem of minimizing $R_\ell(\varphi)$ in (4.4) is, in essence, a factor analysis problem or can be considered as a special case of Karhunen-Loeve expansion [25,26].

Substituting (4.6) into (4.4), we get

$$R_\ell(\varphi) = E\left[\left|\sum_{k=\ell+1}^{N} (\varphi_k^T Z)\,\varphi_k\right|^2\right] = \sum_{k=\ell+1}^{N} \varphi_k^T K_Z \varphi_k \qquad (4.7)$$

where $K_Z = E[ZZ^T]$, is the covariance matrix for X. Minimizing $R_\ell(\varphi)$ is equivalent to minimizing each term in $\sum_{k=\ell+1}^{N} \varphi_k^T K_Z \varphi_k$. In order to minimize $\varphi_k^T K_Z \varphi_k$, introduce a Lagrange multiplier λ_k and form

$$G(\varphi_k) = \varphi_k^T K_Z \varphi_k - \lambda_k(\varphi_k^T \varphi_k - 1) \qquad (4.8)$$

Taking the partial derivative of $G(\varphi_k)$ with respect to φ_k and letting it equal to zero yields the condition

$$K_Z \varphi_k = \lambda_k \varphi_k \qquad (4.9)$$

Equation (4.9) implies that φ_k is an eigenvector of the covariance matrix K_Z and that λ_k is the corresponding eigenvalue. Substituting (4.9) into (4.7) and using the orthonormal condition, we obtain the minimum mean square error as

$$R_\ell(\varphi) = \sum_{k=\ell+1}^{N} \varphi_k^T \lambda_k \varphi_k = \sum_{k=\ell+1}^{N} \lambda_k \qquad (4.10)$$

where λ_k are the $(N-\ell)$ eigenvalue of K_Z corresponding to the eigenvectors φ_k, $k=\ell+1,\ldots,N$. Hence, $R_\ell(\varphi)$ will assume a minimum if we select the first ℓ eigenvectors corresponding to the largest eigenvalues of K_Z.

The new feature space Ω_F will have ℓ dimensions (or features) with coordinates $f_1(X), f_2, \ldots, f_\ell(X)$. The transformation equation is

$$F = \Phi Z \qquad (4.11)$$

where

$$F^T = [f_1(X), f_2(X), \ldots, f_\ell(X)] \quad \text{and} \quad \Phi^T = [\varphi_1^T, \varphi_2^T, \ldots, \varphi_\ell^T] \qquad (4.12)$$

An alternative criterion for feature space transformation is based on the concept of clustering. The idea is to transform the given feature space into a new feature space (with lower dimensions) such that patterns belonging to the same class will be clustered together. Mathematically, this clustering problem can be expressed as a problem of minimizing the mean square intraset distance (distances between patterns belonging to the same class). Let $Z = X - M$, as defined in (4.1). Select arbitrarily a pattern Z_1 from the same class. Form the square of all distances from adjacent patterns Z. The choice of Z_1 is assumed to be independent of the selection of Z. Denote, for a given pattern class ω,

$$p(Z, Z_1/\omega) = p(Z/\omega)\,p(Z_1/\omega) = p p_1 \qquad (4.13)$$

Thus, the mean square intraset distance of class ω is

$$\begin{aligned}\overline{d^2} &= E_{pp_1}\left[|Z - Z_1|^2\right] = E_{pp_1}\left[(Z-Z_1)^T(Z-Z_1)\right] \\ &= E_p\left[Z^T Z\right] + E_{p_1}\left[Z_1^T Z\right] = 2E_p\left[Z^T Z\right] = 2\ \mathrm{tr}\ K_Z\end{aligned} \qquad (4.14)$$

where tr K_Z represents the trace of covariance matrix K_Z.

Let the transformation equation be, as in (4.11), $F = \Phi Z$. Then in the new feature space Ω_F (with dimensionality ℓ) the mean square intraset distance will be

$$\overline{d_F^2(\varphi)} = 2\ \mathrm{tr}\ [\Phi K_Z \Phi^T] \qquad (4.15)$$

Since $K_Z \varphi_k = \lambda_k \varphi_k$, then using the orthonormal condition,

$$\Phi K_Z \Phi^T = \begin{bmatrix} \varphi_1^T \\ \varphi_2^T \\ \vdots \\ \varphi_\ell^T \end{bmatrix} \left[\lambda_1\varphi_1, \lambda_2\varphi_2, \ldots, \lambda_\ell\varphi_\ell\right] = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots & \\ 0 & & & \lambda_\ell \end{bmatrix} \tag{4.16}$$

Substituting (4.16) into (4.15) and carrying out a minimization procedure on $\overline{d_F^2}$ similar to the one used before but with the additional constant volume constraint (by which we mean that the Jacobian of the transformation $F=\Phi Z$ is unity), we obtain the minimum mean square intraset distance as

$$\overline{d_F^2(\varphi)} = 2\ell \left(\prod_{k=1}^{\ell} \lambda_k \right)^{1/\ell} \tag{4.17}$$

Equation (4.17) implies that if we want to choose a set of features such that the transformed pattern vectors have a minimum mean square intraset distance, then the eigenvectors corresponding to the ℓ smallest eigenvalues of covariance matrix K_Z must be chosen.

It is interesting to see the difference between the feature selection procedure developed on the basis of minimizing the mean square intraset distance and the one developed on the basis of minimizing the mean square approximation error. On the other hand, it may not be surprising since the results are developed in terms of two completely different criteria. However, this example does indicate that a good feature selection criterion has to be specified first before a meaningful result can be obtained. It has been suggested that a good feature selection criterion must represent the performance of the overall pattern recognition system. This implies that the design techniques used for feature extraction and selection and for classification are closely related and should not be considered separately. The feature selection techniques introduced in the next section will illustrate this. Nevertheless, because of its simplicity of computation and its independence from the classifier structure chosen, the feature space transformation approach for feature extraction and selection so far has been used in many practical applications.[†]

4.2 Divergence and Separability Measure

It has been pointed out in section 4.1 that the classification accuracy should be considered as a criterion for feature selection. The feature selection techniques described in section 4.1 concern only patterns belonging to the same class. The relations (for example, the separation) between classes have not been considered. In order to use the classification accuracy as a feature selection criterion, an explicit relationship between features and the classification accuracy (for a prespecified classifier structure) needs to be established. Unfortunately, this relationship is not very easy to obtain. Instead, an intermediate quantity, which is related to the classification accuracy or the discriminatory power, is often used as a basis for feature selection. Both the divergence and the average information about the pattern classes characterized by the features have been proposed as criteria for feature selection. The concept of divergence is closely related to the discriminatory power between two pattern classes with normally distributed feature measurement vector X. The use of information measure and divergence as criteria for feature selection or ordering is also implied by the comparison of the expected risks when the Bayes' decision rule is employed for the classification process.

[†] Of course, this may be due to the fact that up to now there have not been many good feature extraction and selection techniques available (Preface of [16]).

The application of divergence as a criterion for feature selection is briefly introduced in the following. Assume that, for ω_i, X is distributed according to a multivariate gaussian density function with mean vector M_i and covariance matrix K. Let the likelihood ratio be

$$\lambda = \frac{p(X/\omega_i)}{p(X/\omega_j)} \tag{4.18}$$

and let

$$L = \log \lambda = \log p(X/\omega_i) - \log p(X/\omega_j) \tag{4.19}$$

For gaussian assumption, we obtain

$$L = X^T K^{-1}(M_i - M_j) - \frac{1}{2}(M_i + M_j)^T K^{-1}(M_i - M_j) \tag{4.20}$$

and

$$E[L/\omega_i] = \frac{1}{2}(M_i - M_j)^T K^{-1}(M_i - M_j) \tag{4.21}$$

Similarly,

$$E[L/\omega_j] = -\frac{1}{2}(M_i - M_j)^T K^{-1}(M_i - M_j) \tag{4.22}$$

and

$$\mathrm{Var}[L/\omega_i] = \mathrm{Var}[L/\omega_j] = (M_i - M_j)^T K^{-1}(M_i - M_j) \tag{4.23}$$

Define the divergence between ω_i and ω_j as

$$J(\omega_i, \omega_j) = E[L/\omega_i] - E[L/\omega_j] \tag{4.24}$$

Then, from (4.21) and (4.22)

$$J(\omega_i, \omega_j) = (M_i - M_j)^T K^{-1}(M_i - M_j) \tag{4.25}$$

It is noted that in (4.23), if K = I, the identity matrix, then $J(\omega_i, \omega_j)$ represents the squared distance between M_i and M_j. If a fixed sample or nonsequential Bayes classification rule is used for the classifier, then for $P(\omega_i) = P(\omega_j) = 1/2$, from (2.8) $X \sim \omega_i$ if $\lambda > 1$, or $L > 0$ or $X \sim \omega_j$ if $\lambda < 1$ or $L < 0$. The probability of misclassification is

$$P(e) = \frac{1}{2} P\{L > 0/\omega_j\} + \frac{1}{2} P\{L < 0/\omega_i\} \tag{4.26}$$

From (4.20)-(4.25), it is concluded that $p(L/\omega_i)$ is a gaussian density function with mean (1/2)J and variance J. Similarly, $p(L/\omega_j)$ is also a gaussian density function with mean (-1/2)J and variance J. Thus,

$$P(e) = \frac{1}{2}\int_0^{\infty} \frac{1}{\sqrt{2\pi J}} \exp[-\frac{1}{2}(\xi + \frac{1}{2}J)^2/J]d\xi + \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi J}} \exp[-\frac{1}{2}(\xi - \frac{1}{2}J)^2/J]d\xi \tag{4.27}$$

Let

$$y = \frac{\xi \pm \frac{1}{2}J}{\sqrt{J}} \tag{4.28}$$

then

$$P(e) = \int_{\frac{1}{2}\sqrt{J}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}y^2] dy \tag{4.29}$$

It is noted from (4.29) that, P(e) is a monotonically decreasing function of $J(\omega_i, \omega_j)$. Therefore, features selected or ordered according to the magnitude of $J(\omega_i, \omega_j)$ will imply their corresponding discriminatory power between ω_i and ω_j.

For more than two pattern classes, the criterion of maximizing the minimum divergence or the expected divergence between any pair of classes has been proposed. In the case that the covariance matrices K_i and K_j are not equal and the number of classes is more than two, the following approach has been proposed [15]. Let $D_i(X)$ be the discriminant function of class ω_i and class ω_j can be expressed as

$$P_{ij}(e) = \frac{1}{2} P\left\{D_i(X) > D_j(X)/\omega_j\right\} + \frac{1}{2} P\left\{D_i(X) < D_j(X)/\omega_i\right\} \tag{4.30}$$

with the assumption of equal a priori probabilities. Suppose that a family of linear discriminant functions is chosen for $D_i(X)$ as

$$D_i(X) = B_i^T X - c_i, \qquad i=1,\dots,m \tag{4.31}$$

where B_i is an N-dimensional vector and c_i is the constant term. Let

$$D_i(X) - D_j(X) = (B_i - B_j)^T X - (c_i - c_j) = B_{ij}^T X - c_{ij} \tag{4.32}$$

Then (4.30) becomes

$$P_{ij}(e) = \frac{1}{2} P\left\{B_{ij}^T X > c_{ij}/\omega_j\right\} + \frac{1}{2} P\left\{B_{ij}^T X < c_{ij}/\omega_i\right\} \tag{4.33}$$

After defining a standardized gaussian random variable ξ, (4.33) can be written as

$$P_{ij}(e) = \frac{1}{2}\{2 - P[\xi < d_j] - P[\xi < d_i]\} \tag{4.34}$$

where

$$d_i = \frac{B_{ij}^T M_i - c_{ij}}{(B_{ij}^T K_i B_{ij})^{1/2}} \quad (4.35) \qquad \text{and} \qquad d_j = \frac{c_{ij} - B_{ij}^T M_j}{(B_{ij}^T K_j B_{ij})^{1/2}} \tag{4.36}$$

Since, in this case, for linear classification procedures, the minimax procedure is admissible, it is used here to determine an appropriate criterion for feature selection. From (4.34), the condition of minimizing the maximum of $P_{ij}(e)$ is

$$d_i = d_j = d_{ij} \tag{4.37}$$

Thus, it can be obtained from (4.35), (4.36) and (4.37) that

$$d_{ij} = \frac{B_{ij}^T (M_i - M_j)}{(B_{ij}^T K_i B_{ij})^{1/2} + (B_{ij}^T K_j B_{ij})^{1/2}} \tag{4.38}$$

The value of B_{ij}, which maximizes d_{ij} is of the form

$$B_{ij} = [\lambda_{ij} K_i + (1-\lambda_{ij}) K_j]^{-1} (M_i - M_j) \tag{4.39}$$

where λ_{ij} is a Lagrange multiplier and can be calculated by solving the equation

$$B_{ij}^T \, [\lambda_{ij}^2 K_i - (1-\lambda_{ij})^2 K_j] \, B_{ij} = 0 \tag{4.40}$$

with $0 < \lambda_{ij} < 1$. Finally, (4.34) becomes

$$P_{ij}^*(e) = 1 - P[\xi < d_{ij}] \tag{4.41}$$

It is noted from (4.41) that a monotonic functional relationship exists between $P_{ij}^*(e)$ and d_{ij}. The quantity d_{ij} is denoted as a separability measure between class ω_i and class ω_j, and the divergence $J(\omega_i, \omega_j)$ can be considered as a special case of d_{ij} [15].

For multiclass classification problems, the performance of classification can be measured in terms of a weighted sum of all pairwise misclassifications; i.e., the overall performance measure is

$$P(e) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{m} P_{ij}(e), \qquad i \neq j \tag{4.42}$$

Based on the separability measure d_{ij}, the feature selection criterion suggested is to maximize the expected separability measure for all pairs of classes; i.e.,

$$\text{Max}\left\{\sum_{i=1}^{m} \sum_{j=1}^{m} P(\omega_i)\, P(\omega_j)\, d_{ij}\right\} \qquad i \neq j \tag{4.43}$$

In practical applications, the monotonic properties demonstrated by (4.29) and (4.41) sometimes cannot be obtained; i.e., it is possible to have subsets of features resulting in better classification performance than that using the complete set of features [7]. This is probably due mainly to the deviation of the actual feature distributions from the gaussian assumption and/or the error involved in the estimation of parameters M_i and K_i using an insufficient number of training samples.[†]

4.3 Illustrative Examples

The feature extraction and selection techniques discussed in Sections 4.1 and 4.2 have been applied to crop classification problems. Five major crops were considered: soybeans, corn, oats, wheat and red clover. Gaussian assumptions were used for each crop class with mean vectors and covariance matrices estimated from approximately 400 training samples per class. Equal a priori probabilities were assumed. The maximum likelihood classification rule was used to empirically determine the classification accuracy.

Example 3: Feature Space Transformation for Minimizing the Mean Square Approximation Error. The following computational procedure was used.

Step 1: Compute $Z = X-M$ and K_Z for each class from sample mean vectors and sample covariance matrices.

Step 2: Find the eigenvalues and their corresponding eigenvectors for K_Z. Let the eigenvectors be lexicographically arranged according to the decending order of their corresponding eigenvalues. Choose the first ℓ eigenvectors.

Step 3: Make the transformation defined in (4.11). The new feature vector will be an ℓ-dimensional vector F. Results of this experiment are shown in Table 4.

Example 4: Feature Selection Using Separability Measure. The separability measure defined in (4.38) was applied to the feature selection in crop classification problem. For the five-class problem under consideration, (4.43) was used. Experimental results are given in Table 5.

V. CONCLUSIONS AND REMARKS

In this paper, several major pattern recognition techniques are introduced to demonstrate their applicability to data interpretation problems in remote sensing systems. Illustrative examples are based on crop classification results. In order to keep the paper to a reasonable size, only those techniques which have been applied to real data experiments are presented. Undoubtedly there are useful or potentially useful techniques for solving remote sensing problems which have not been included. Methods relating to data interpretation directly through picture or image processing, as mentioned in Section I, are considered to be beyond the scope of this paper.

Parametric descriptions of pattern (vector) characteristics for each class are the required a priori information in applying parametric classification techniques. In the absence of such a priori information one has to either use nonparametric classification techniques or estimate (or learn) the required information by applying appropriate estimation methods. The estimated information will then be used for classification purposes. However, in order to have good estimation accuracy, a large number of learning samples is usually required. Clustering and mode estimation techniques, though not covered in this paper, are considered quite effective in learning the data structure, particularly when performed in a man-machine interactive scheme [27-31]. Perhaps due to the lack of a solid theoretical treatment, a designer's experience and preference have often played an important role in applying clustering techniques. An important problem, but unfortunately with much less fruitful results compared with the classification problem, is feature extraction and selection. Feature space (linear) transformation techniques have been commonly used (although not always satisfactorily) in reducing a large number of features into a relatively smaller number because they are simple to implement and are nonparametric in nature. In addition to divergence and separability measure, other similar quantities such as the entropy and other distances (for example, Bhattacharyya distance) have also been suggested as feature selection criteria in parametric cases. However, except for a few special cases, simple and explicit relationships cannot be established between the features used and the performance of the recognition system. In many practical applications, project objectives, the designer's intuition and his experience about the data under study are still the major factors in determining the features to be used.

[†] A rule of thumb commonly used in practice is that the training samples required should be at least 10 N, where N is the number of features used.

Practical applications such as remote-sensing data interpretations, usually provide challenging problems in pattern recognition research. On the other hand, some fruitful results have already been obtained in the application of pattern recognition methods to the data interpretation problems in remote sensing systems. It is hoped that this paper will stimulate more interest in research and more exchanges of ideas between these two areas--pattern recognition and remote sensing.

REFERENCES

1. K. S. Fu, Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.
2. N. J. Nilsson, Learning Machines, McGraw-Hill, New York, 1965.
3. R. O. Duda, "Elements of Pattern Recognition," Chap. 1 in Adaptive, Learning and Pattern Recognition Systems: Theory and Applications, ed. by J. M. Mendel and K. S. Fu, Academic Press, New York, 1970.
4. K. S. Fu, "Statistical Pattern Recognition," Chap. 2 in Adaptive, Learning and Pattern Recognition Systems: Theory and Applications, ed. by J. M. Mendel and K. S. Fu, Academic Press, New York, 1970.
5. G. Nagy, "State of the Art in Pattern Recognition," Proc. IEEE, Vol. 56, No. 5, May 1968.
6. K. S. Fu, "On the Preliminary Progress in Crop Classification Research," IEEE Intl. Pattern Recognition Workshop, Aug. 12-16, 1968, Delft, Holland.
7. K. S. Fu, D. A. Landgrebe and T. L. Phillips, "Information Processing of Remotely Sensed Agricultural Data," Proc. IEEE, Vol. 57, No. 4, April 1969, pp. 639-653.
8. G. C. Cheng, ed. Pictorial Pattern Recognition, Thompson Book Co., Washington, D. C., 1968.
9. A. Rosenfeld, Picture Processing by Computer, Academic Press, New York, 1969.
10. A. Grasselli, Automatic Interpretation and Classification of Images, Academic Press, New York, 1969.
11. H. C. Andrews, Computer Techniques in Image Processing, Academic Press, New York, 1970.
12. A. R. Shulman, Optical Data Processing, Wiley, New York, 1970.
13. J. T. Tou and R. P. Heydorn, "Some Approaches to Optimum Feature Extraction," in Computer and Information Sciences--II, ed. by J. R. Rou, Academic Press, New York, 1967.
14. M. D. Levine, "Feature Extraction: A Survey," Proc. IEEE, Vol. 57, No. 8, Aug. 1969, pp. 1391-1407.
15. K. S. Fu, P. J. Min and T. J. Li, "Feature Selection in Pattern Recognition," IEEE Trans. on SSC, Vol. SSC-6, No. 1, Jan. 1970.
16. Special Issue of IEEE Computers Transactions on Feature Extraction and Selection in Pattern Recognition, Vol. C-20, No. 9, September 1971.
17. R. Newman and B. Reisine, "Practical Applications of Sequential Pattern Recognition Techniques," Proc. Seventh Space Congress, April 22-24, 1970, Cocoa Beach, Florida.

18. T. Huang, "Per-Field Classification of Remotely Sensed Agricultural Data," Proc. 8th Allerton Conf. on Circuit and System Theory, Oct. 7-9, 1970, Urbana, Ill., pp. 427-434.
19. F. Rosenblatt, Principles of Neurodynamics, Sparten Books, Washington, D.C., 1961.
20. D. R. Specht, "Generation of Polynomial Discriminant Functions for Pattern Recognition," IEEE Trans. on Elec. Comp., Vol. 16, No. 3, June 1967, pp. 308-319.
21. P. H. Swain and K. S. Fu, "On the Application of a Nonparametric Technique to Crop Classification Problem," Proc. 1968 Natl. Elec. Conf.
22. M. A. Aiserman, E. M. Braverman and L. I. Rozonoer, "The Probability Problem of Pattern Recognition Learning and the Method of Potential Functions," Automatika i Telemekhamika, Vol. 25, 1964, pp. 1307-1323.
23. R. R. Lemke and K. S. Fu, "On the Application of the Potential Function Method to Pattern Classification," Proc. 1968 Natl. Elec. Conf. pp. 107-112.
24. T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. on Infor. Theory, Vol. IT-13, No. 1, Jan. 1967.
25. M. S. Watanabe, "Karhunen-Loeve Expansion and Factor Analysis--Theoretical Remarks and Applications," Proc. 4th Prague Conf. on Infor. Theory, 1965.
26. Y. T. Chien and K. S. Fu, "Selection and Ordering of Feature Observations in a Pattern Recognition System," Information and Control, Vol. 12, No. 5, 1968.
27. G. H. Ball, "Data Analysis in the Social Sciences," Proc. 1965 Fall Joint Computer Conf., Spartan Books, Washington, D. C.
28. G. H. Ball, "Classification Analysis," Technical Note, Project 5533, Stanford Research Institute, Menlo Park, Calif., Nov. 1970.
29. R. M. Haralick and I. Dinstein, "An Iterative Clustering Procedure," Proc. 9th IEEE Symp. on Adaptive Processes, Dec. 7-9, 1970, Austin, Tex.
30. A. G. Wacker and D. A. Landgrebe, "Boundaries in Multispectral Imaging by Clustering," Proc. 9th IEEE Symp. on Adaptive Processes, Dec. 7-9, 1970, Austin, Tex.
31. E. G. Henrichon and K. S. Fu, "On Mode Estimation in Pattern Recognition," Proc. 7th Symp. on Adaptive Processes, Dec. 16-18, 1968, Los Angeles, Cal.
32. R. E. Marshall and F. J. Kriegler, "Multispectral Recognition Techniques: Present Status and A Planned Hybrid Recognition System," Proc. 9th IEEE Symp. on Adaptive Processes, Dec. 7-9, 1970, Austin, Tex.
33. R. D. Joseph and S. S. Viglione, "A Hybrid Computation System for Pattern Classification," Symp. on Automatic Photointerpretation and Recognition, Dec. 1-2, 1970, Baltimore, Md.
34. W. A. Fischer, "Potential Applications of Pattern Recognition Techniques in the EROS Program," Symp. on Automatic Photointerpretation and Recognition, Dec. 1-2, 1970, Baltimore, Md.
35. R. E. Kingly and J. E. Walker, "The Use of Correlated Spectral Imagery in Two Bands for Extraction of Agricultural Information," Symp. on Automatic Photointerpretation and Recognition, Dec. 1-2, 1970, Baltimore.
36. D. Steiner, K. Baumberger and H. Maurer, "Computer Processing and Classification of Multi-Variate Information from Remote Sensing Imagery," Proc. 6th Symp. on Remote Sensing of Environment, Ann Arbor, Mich., 1969.
37. Special Issue of IEEE Proc. on Remote Environmental Sensing, Vol. 57, No. 4, April 1969.
38. K. S. Fu, "On the Application of Pattern Recognition Techniques to Remote Sensing Problems," Tech. Rept. TR-EE 71-13, School of Elec. Engr., Purdue Univ., June 1971.

ACKNOWLEDGMENT

Crop classification experiments were performed at LARS, Purdue University. Discussions with the members of LARS Data Processing Group were definitely helpful.

TABLE 1
CLASSIFICATION SUMMARY BY TRAINING CLASSES

Class	# of sampl.	Percent Correct	Number of Samples Classified Into 1	2	3	4	5	6	7	8	Thresh-old
1(Wheat)	702	99.4	698	0	0	1	0	0	3	0	0
2(Soybeans)	426	98.8	0	421	4	0	0	0	0	1	0
3(Corn)	423	97.9	0	8	414	1	0	0	0	0	0
4(Oats)	423	99.5	0	0	1	421	1	0	0	0	0
5(Clover)	423	98.8	0	0	0	1	418	4	0	0	0
6(Alfalfa)	259	98.8	0	0	0	1	2	256	0	0	0
7(Rye)	330	98.5	4	0	0	1	0	0	325	0	0
8(Bare Soil)	190	100.0	0	0	0	0	0	0	0	190	0
Total	3176		702	429	419	426	421	260	328	191	0

Overall performance = 99.0 Features used: all 12
Average performance by class = 99.0

TABLE 2
CLASSIFICATION SUMMARY BY TEST CLASSES

Class	# of sampl.	Percent Correct	Number of Samples Classified Into 1	2	3	4	5	6	7	8	Thresh-old
1(Soybeans)	7171	95.7	6860	126	0	35	10	1	7	1	131
2(Corn)	2775	91.4	126	2535	0	24	0	73	5	0	12
3(Wheat)	2667	99.2	0	0	2645	12	6	0	0	0	4
4(Oats)	1595	91.7	12	7	18	1463	3	69	20	0	3
5(Rye)	621	97.4	0	0	14	2	605	0	0	0	0
6(Clover)	2333	91.0	21	26	3	43	0	2124	0	0	14
7(Alfalfa)	912	87.9	0	6	0	47	0	47	804	0	10
8(Bare Soil)	332	98.8	4	0	0	0	0	0	0	328	0
Total	18406		7023	2700	2680	1626	624	2314	836	329	174

Overall performance = 94.3 Features used: all 12
Average performance by class = 94.1

TABLE 3
CLASSIFICATION RESULTS OF AIRCRAFT DATA TAKEN FROM AREA C-4 (JULY 1966)

Class	Fields	No. of Fields Classified Into 1	2	3	4
1 (Soybeans)	19	17	2	0	0
2 (Corn)	31	1	30	0	0
3 (Pasture)	18	0	0	18	0
4 (Stubble)	13	0	0	3	10

Total No. Fields Classified = 81
No. Fields Correctly Classified = 75
% Correct Classification of Per-Field Classifier = 92.6
% Correct Classification of MLDR Classifier = 79.7

TABLE 4
RESULTS OF FEATURE SELECTION USING FEATURE SPACE TRANSFORMATION

# of Feat.	Best Feature Set	MLDR % Correct Class.
1	x_9	48.2
2	x_1, x_9	75.6
3	x_1, x_{10}, x_{11}	92.8
4	x_1, x_6, x_{10}, x_{11}	93.8
5	$x_1, x_6, x_9, x_{10}, x_{11}$	94.2

No. of test samples = 7530

TABLE 5
RESULTS OF FEATURE SELECTION USING
SEPARABILITY MEASURE CRITERION

Number of Features	Best Feature Set	d_t	MLDR % Correct Classification
1	x_9	17.9	62.4
2	x_1, x_9	29.7	89.4
3	x_1, x_{10}, x_{11}	33.8	95.0
4	x_1, x_6, x_{10}, x_{11}	35.1	95.1
5	$x_1, x_6, x_9, x_{10}, x_{11}$	36.2	95.6
6	$x_1, x_6, x_9, x_{10}, x_{11}, x_{12}$	36.8	95.8

$d_t = \sum_{ij} d_{ij}$ Number of test samples = 7530

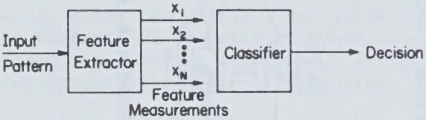

X_1
X_2
X_N
Input
Pattern
Feature
Extractor
Classifier
Decision
Feature
Measurements

FIGURE 1. A PATTERN RECOGNITION SYSTEM.

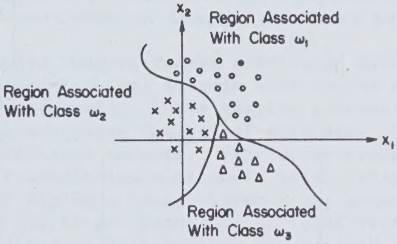

X_2
Region Associated
With Class ω_1
Region Associated
With Class ω_2
X_1
Region Associated
With Class ω_3

FIGURE 2. AN EXAMPLE OF PARTITION IN A TWO DIMENSIONAL FEATURE SPACE.

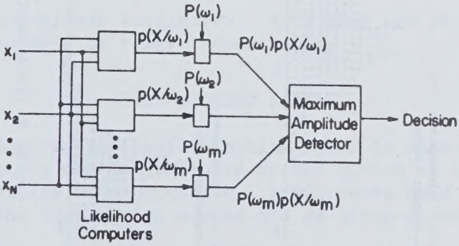

$P(\omega_1)$
$p(X/\omega_1)$
$P(\omega_1)p(X/\omega_1)$
X_1
$P(\omega_2)$
$p(X/\omega_2)$
X_2
Maximum
Amplitude
Detector
Decision
$P(\omega_m)$
$p(X/\omega_m)$
X_N
$P(\omega_m)p(X/\omega_m)$
Likelihood
Computers

FIGURE 3. A BAYES CLASSIFIER.

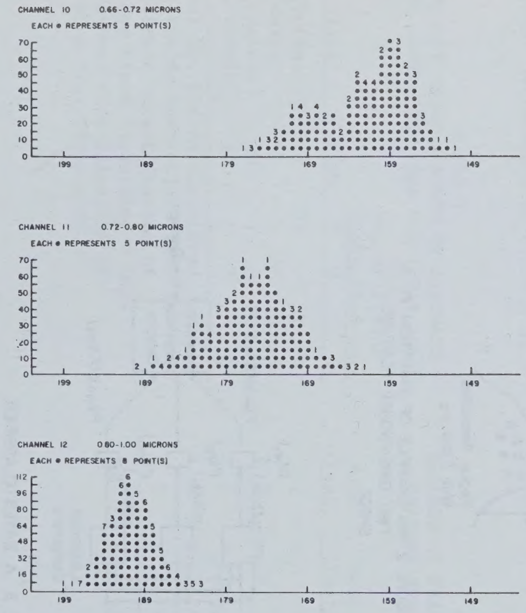


FIGURE 4. TYPICAL HISTOGRAMS FOR WHEAT

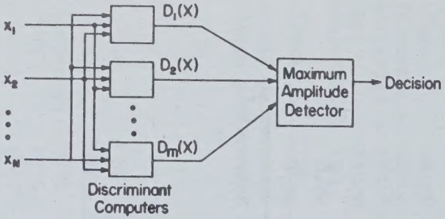

X_1
X_2
X_N
$D_1(X)$
$D_2(X)$
$D_m(X)$
Discriminant Computers
Maximum Amplitude Detector
Decision

FIGURE 5. A CLASSIFIER.

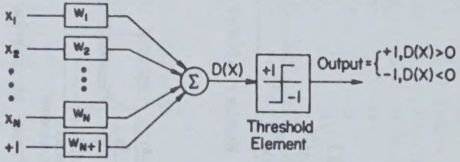

X_1
X_2
X_N
+1
W_1
W_2
W_N
W_{N+1}
Σ
D(X)
+1
-1
Threshold Element
$\text{Output}=\begin{cases}+1, D(X)>0\\-1, D(X)<0\end{cases}$

FIGURE 6. A LINEAR TWO-CLASS CLASSIFIER.

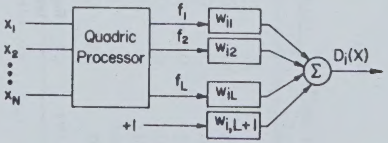

X_1
X_2
X_N
Quadric Processor
f_1
f_2
f_L
+1
W_{i1}
W_{i2}
W_{iL}
$W_{i,L+1}$
Σ
$D_i(X)$

FIGURE 7. A QUADRIC DISCRIMINANT COMPUTER.

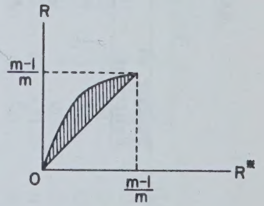


FIGURE 8. BOUNDS ON THE NN RISK.

IMAGING RADAR TECHNIQUES FOR REMOTE SENSING APPLICATIONS

JERRY S. ZELENKA
The University of Michigan
Willow Run Laboratories
Ann Arbor, Michigan

ABSTRACT

The basic concepts of fine-resolution, imaging radar systems are reviewed. Both side-looking and hologram (downward-looking) radars are described and compared. Several examples of microwave imagery obtained with these two types of systems are shown.

INTRODUCTION

Synthetic-aperture radar [1-5] is one of the most sophisticated electronic systems to be developed during the 1960's. This particular system development is the result of extensive investigations in range-Doppler radars, coherent optical recorders and processors, aircraft motion compensation, automatic shape recognition, and large dynamic range displays. Although most of the theoretical concepts of synthetic-aperture techniques are now well understood, engineering development in the areas of data reduction and parameter optimization is still required.

The purpose of this paper is twofold. First, a brief description of two microwave imaging systems which utilize synthetic-aperture techniques is presented. Second, some preliminary experimental results obtained at X- and K_a-bands (i.e., 3 and 1.78 cm, respectively) are given to illustrate the potential usefulness of imaging microwave systems to remote sensing.

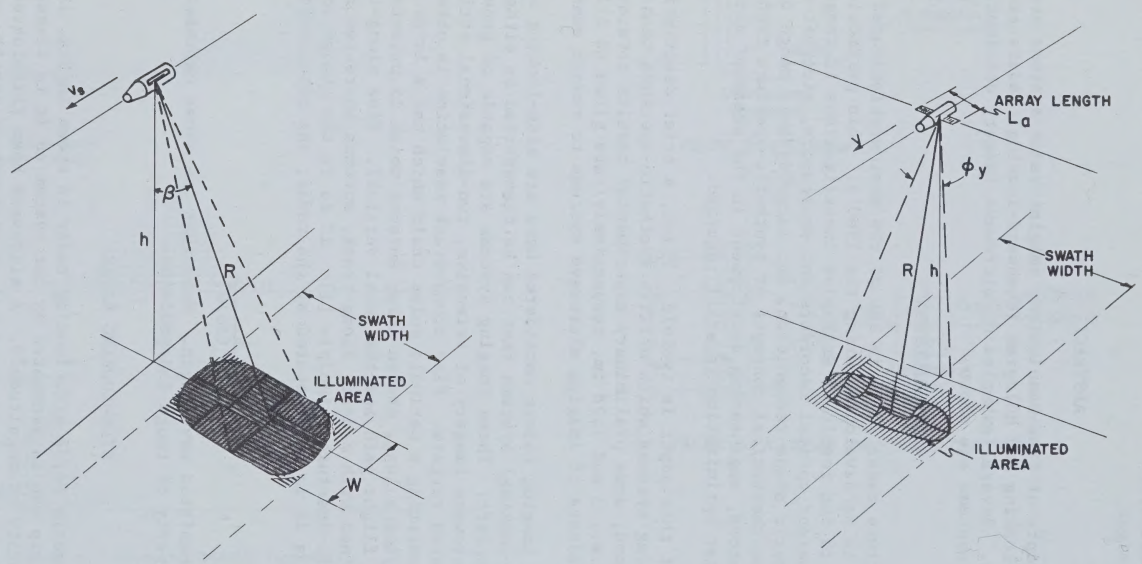
The microwave imaging radars considered here are side-looking and hologram (downward-looking) systems that can be transported by either an aircraft or a spacecraft. These imaging systems are capable of generating fine-resolution microwave imagery of extensive, two-dimensional strips of terrain and associated targets. Fine cross-track resolution is obtained either (1) by transmitting a periodic pulse train which has a large bandwidth or (2) by flying a long, phased-array antenna which is oriented normal to both the flight path and the local vertical. Fine along-track resolution is obtained with modest, along-track, antenna apertures by the use of the synthetic-aperture principle [1]: if $\Delta\alpha$ is the change in aspect angle of a target as it is interrogated by the radar, the obtainable resolution is

$$\rho = \frac{\lambda}{4 \sin (\Delta\alpha/2)} \qquad (1)$$

where λ is the transmitted wavelength. Note that the above expression has the remarkable property of range independence.

SIDE-LOOKING RADAR

A typical geometry for a side-looking radar is shown in Fig. 1A. A two-dimensional strip map is generated by the system as it is transported at a constant velocity of magnitude v. A microwave beam illuminates the entire width of the swath being mapped and an along-track portion of approximately



v_s
h
β
R
SWATH WIDTH
ILLUMINATED AREA
W
ARRAY LENGTH
L_a
γ
ϕ_y
R
h
SWATH WIDTH
ILLUMINATED AREA

A) SIDE-LOOKING MICROWAVE HOLOGRAM RADAR SPACECRAFT GEOMETRY

B) MICROWAVE HOLOGRAM RADAR SPACECRAFT GEOMETRY

FIGURE 1. IMAGING RADAR GEOMETRIES

$$W = \frac{\lambda}{d}R \quad (2)$$

where R is the slant range to the target area and d the effective aperture length of the antenna.

The radar echoes are recorded for subsequent Doppler and possible range processing. Currently, this is usually done with an optical recorder consisting primarily of a CRT, relay lens, film, and associated transport. After the returns from a transmitted pulse to video are coherently mixed, the resulting signal intensity modulates a CRT about a constant bias level, while an image of its spot is swept across a photographic film at a constant velocity. Therefore, there is a linear correspondence between slant range (time-delay) and the spatial dimension across the film. The returns from successive pulses are recorded side by side along the film by transporting the film past the CRT at a rate proportional to the radar speed v. Thus, the spatial dimension along the film is linearly related to the along-track dimension of the terrain. After the data film is developed, it is usually referred to as a "signal film."

The radar data is stored on a signal film and transformed into an optical image by means of a coherent optical processor or correlator. Such a processor can be regarded as an analog computer that reconstructs in the visible spectrum a scaled-down replica of the original microwave wavefront that was intercepted by the radar system in the along-track dimension. A series of lenses and stops are then used to reconstitute a visible, scaled-down image of the original scene.

If short pulses are transmitted by the radar system, the range image occurs right on the signal film, and it merely has to be reimaged at the output of the processor. If a coded pulse is used, a matched filter must be incorporated into the range dimension of the processor. Fortunately, for the popular, linear, FM (chirp) waveform, the range image is automatically formed near the data film by illumination of the film with collimated laser (monochromatic) light. Then the processor simply brings the range- and azimuth-image planes into coincidence. The optical processor is also used to change range- and along-track scaling factors and to perform special filtering operations. The output of the optical processor can be viewed directly or recorded on another film called the output film.

The along-track resolution, theoretically achievable with a synthetic-aperture system, is about d/2. This result is obtained by approximating the along-track beamwidth by

$$\Delta\alpha = \frac{\lambda}{d} \quad (3)$$

Then, by substituting Eq. (3) into (1) and assuming that $\Delta\alpha << 1$, we get the desired result

$$\rho_a = \frac{d}{2} \quad (4)$$

The reason that ρ_a is independent of range R is that the length of the synthetic aperture which can be synthesized is proportional to R; thus, the F-number of the system is independent of range. In any practical system, the achievable, along-track resolution is usually considerably larger than d/2, because uncompensated motion errors prohibit diffraction-limited performance over the entire along-track history.

Fine cross-track resolution is usually obtained by transmitting a period-pulse train which has a large bandwidth. The resulting slant-range resolution for a short-pulse system is

$$\rho_R = \frac{c\tau}{2} \approx \frac{c}{2\ \Delta f} \tag{5}$$

where τ is the transmitted pulse length, c is the velocity of light in free space, and Δf is the required radio frequency (RF) bandwidth. If long, coded pulses are transmitted and the resulting returns are match-filtered, the slant range resolution is approximately [1,2,6]

$$\rho_R = \frac{c}{2\ \Delta f} \tag{6}$$

From the geometry, the cross-track (ground-range) resolution is

$$\rho_c = \rho_R \csc \beta = \frac{c}{2\ \Delta f} \csc \beta \tag{7}$$

where β is the angle between the line of sight to the target and the local vertical.

HOLOGRAM RADAR

The geometry for a hologram radar is shown in Fig. 1B. As with side-looking radar, hologram radar is assumed to be transported at a constant velocity of magnitude v. Note that the region to be imaged is now a strip directly beneath the flight path rather than out to the side. A constant-frequency, microwave source illuminates a portion of the target field as shown in Fig. 1B. The wavefront scattered back to the hologram radar is intercepted by a long linear, phased-array antenna oriented normal to the flight path and the local vertical.

Each element in the array is followed by a receiver and coherent detector in order to determine the amplitude and phase resulting along the array. This amplitude and phase information is linearly mapped across a strip of photographic film with the use of a multiplexer and an optical recorder. To produce a two-dimensional, optical hologram which is a recording of the scaled-down, microwave wavefront intercepted by the transported linear array, synthetic-aperture techniques are applied to each channel of the array. In other words, the amplitude and phase history experienced by each channel of the array is recorded along the signal film or hologram.

As in conventional optical holography [7-9], the scaled-down microwave hologram which results can be illuminated with a monochromatic, spatially coherent light source to form a scaled-down image of the original scene. However, an optical processor is usually used for spatial filtering and scale changes.

The along-track resolution capability is essentially the same as for side-looking systems, since the same principles are involved. Thus

$$\rho_a = \frac{\lambda}{4 \sin (\Delta\alpha/2)} \approx \frac{d}{2} \tag{8}$$

However, uncompensated motion errors often prohibit us from achieving the above theoretical result. The cross-track resolution corresponds to conventional, imaging aperture theory; that is, for an array cross-track dimension of effective length L_a

$$\rho_c = \frac{\lambda}{L_a} R\ (\sec\ \beta)^2 = \frac{\lambda}{L_a} h\ (\sec\ \beta)^3$$

where h is the altitude of the radar.

SIDE-LOOKING VERSUS HOLOGRAM SYSTEM

In addition to the downward-looking capability of a hologram radar system, there are other pronounced distinctions between it and a side-looking system. A hologram radar requires no pulse modulation to achieve two-dimensional imaging. Consequently, its transmitter can produce a high average power without the peak power limitation imposed on pulsed systems. A hologram radar system generates images which have angle-angle coordinates rather than slant range-angle. Such microwave imagery is especially desirable if optical or infrared imaging systems are also utilized, since all these systems have the same format.

A hologram radar system can yield high-quality, three-dimensional information by simultaneously transmitting more than one constant frequency. If two frequencies are transmitted, interference fringes, each of which corresponds to targets at a fixed slant range, are superimposed on the resulting image. The change in slant range between successive contours is simply

$$\delta_R = \frac{c}{2\ \Delta f} \tag{9}$$

where Δf as used here is the separation between the two frequencies. To obtain three-dimensional information from a side-looking radar system, either triangulation techniques (i.e., measuring the lengths of shadows) must be used or multiple passes by the target area--possibly with an interferometric antenna--must be exercised.

Undoubtedly, the greatest difference between hologram and side-looking radar systems is that the cross-range resolution obtainable with a hologram radar is proportional to range, while for a side-looking radar, it is range-independent. For this reason, long-range, fine resolution microwave imaging systems (e.g., from a spacecraft) must be, for the most part, side-looking systems. One possible exception is a short-wavelength system (e.g., a few millimeters) which uses a very long array (i.e., 10 m or more). Since the cross-track resolution of a hologram radar is proportional to altitude, these short-wavelength systems have a built-in zoom capability if altitude can be varied.

Although side-looking radar systems require large bandwidth networks in order to achieve fine slant-range or cross-range resolution, they are generally considered to be much simpler to construct. A side-looking radar system requires only one antenna element and one receiver, while a hologram radar should have at least 100 array and receiver elements, possibly as many as a few thousand. For a side-looking system, the required compensation for spurious motions in the radar vehicle is considerably simpler, since there is only one receiving element to consider. A long receiving array in a hologram system may also experience flexing caused by abrupt vehicular motions. Although cumbersome, these motions must also be compensated to realize diffraction-limited performance.

EXPERIMENTAL RESULTS

To date, only one hologram radar system has ever been built.* This demonstration system has a cross-range angular resolution of 10.7°, which corresponds to 5.4 m at an altitude of 300 m. Its along-track resolution is about 0.3 m. This system can map cross-range swaths corresponding to ±45°, which is about 600 m at its nominal altitude of 300 m. Operating this radar in its dual-frequency mode, 30 m range contour intervals are superimposed on its output map.

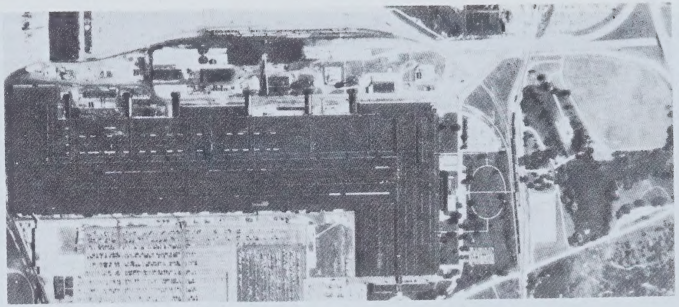
Two examples of imaging produced with the hologram radar system and corresponding aerial photographs are shown in Figs. 2 and 3. Figure 2B is an example of the single-frequency mode of operation. The outline and superstructure of the GM building near Willow Run Airport are very distinct. An example of dual-frequency operation is shown in Fig. 3B. The slight displacement in one range contour as it passed over the building was used together with the radar geometry to measure the height of the one-story building to an accuracy of a few feet. The various contours in the vicinity of the gravel pit accurately depict the ruggedness of the terrain.

Numerous side-looking radars have been constructed. Most of these systems operate at wavelengths from L-band (0.5 m) to K_u-band (1 cm). However, systems are being designed and constructed over the spectrum ranging from HF (60 m) to E-band (3 mm). Modifications have just been completed on one of The University of Michigan's X-band (3 cm) side-looking radars so that it can also have an L-band (25 cm) capability. This modified system is capable of receiving both parallel and cross polarizations. A few examples of X-band imagery generated by this system are shown in Fig. 4. All of this imagery has been specially processed for 50' x 50' resolution, although finer resolution is achievable with this system.

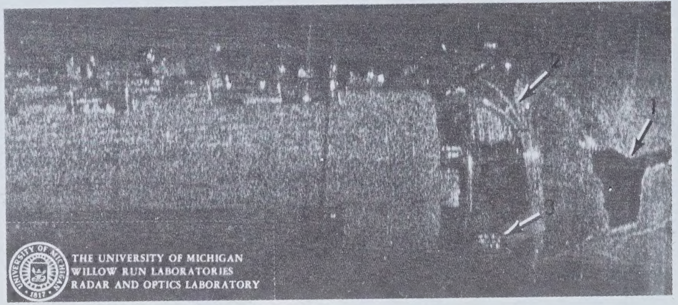
Figure 4 presents four examples of X-band imagery obtained with the use of parallel polarization. In Fig. 4A, the major cracks in the ice covering Lake Superior are clearly visible, although a light layer of snow prevents most of them from being photographed. The ice breaker and bulk carrier making their way off Whitefish Point are relatively easy to spot. The prehistoric coast line at Whitefish Point is quite conspicuous in Fig. 4B because of the drainage patterns which have produced shallow ridges. The breakwater and radio beacon are also detectable. Several examples of cultural (man-made) targets are shown in Figs. 4C and 4D. In Fig. 4C, the airport runways and associated structures are very conspicuous. However, closer examination of the figure reveals other targets such as a row of street lights and fixtures and a set of instrument approach lights. Figure 4D consists of a rural area in norther Michigan. The indicated drainage pattern has been mapped with considerable detail. The individual fields show up in a checkerboard fashion since the radar cross-section density varies from field to field. Clumps of trees and their associated shadows can also be identified.

---

*A 100-element hologram radar system operating at 1.78 cm wavelength has been constructed and flight tested by the Radar and Optics Division of Willow Run Laboratories of The University of Michigan.



A) Aerial Photograph of GM Building


1
2
3
THE UNIVERSITY OF MICHIGAN
WILLOW RUN LABORATORIES
RADAR AND OPTICS LABORATORY

B) Radar Image of GM Building
1) Small Lake
2) Service Roads
3) Parked Vehicles

FIGURE 2. HOLOGRAM RADAR IMAGERY (SINGLE-FREQUENCY OPERATION)

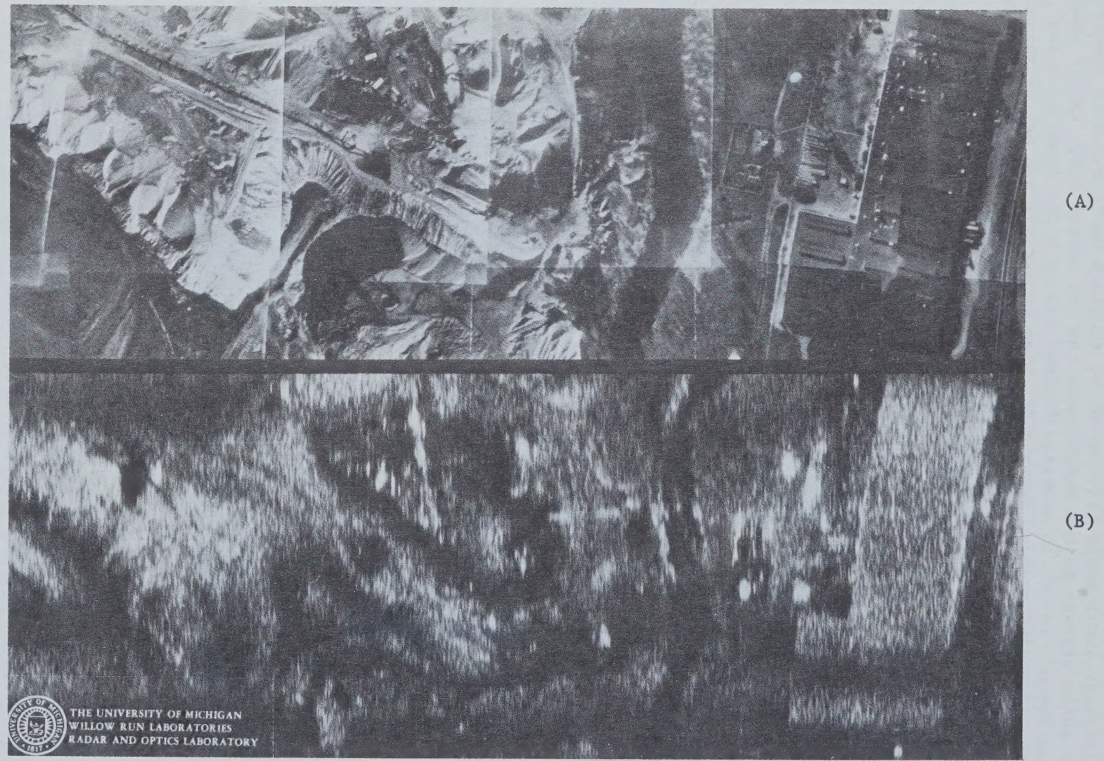
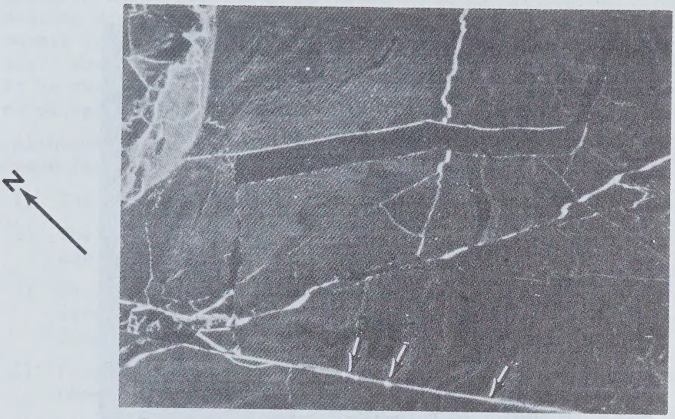
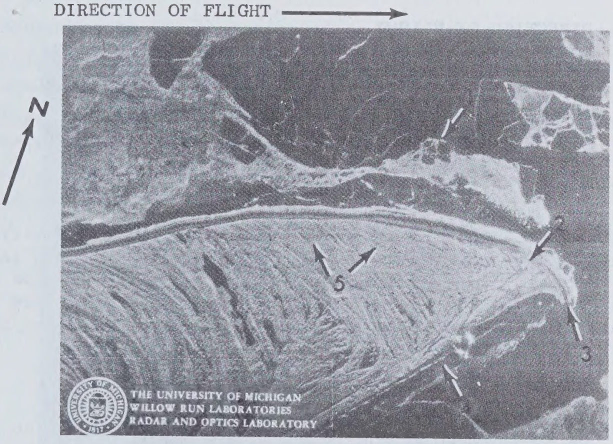

(A)
(B)
THE UNIVERSITY OF MICHIGAN
WILLOW RUN LABORATORIES
RADAR AND OPTICS LABORATORY

FIGURE 3. HOLOGRAM RADAR IMAGERY (DUAL-FREQUENCY OPERATION)
A) Aerial Photograph of King-Seeley Plant and Gravel Pit
B) Radar Image of King-Seeley Plant and Gravel Pit

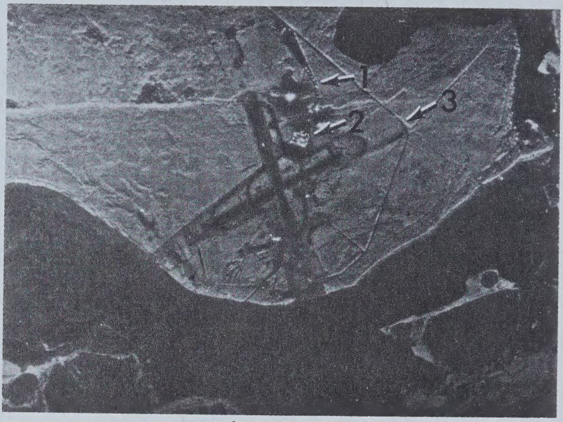

N
DIRECTION OF FLIGHT

A) ICE STRUCTURE WHITEFISH BAY
1) Shipping Channel
2) Ice Breaker
3) Bulk Carrier

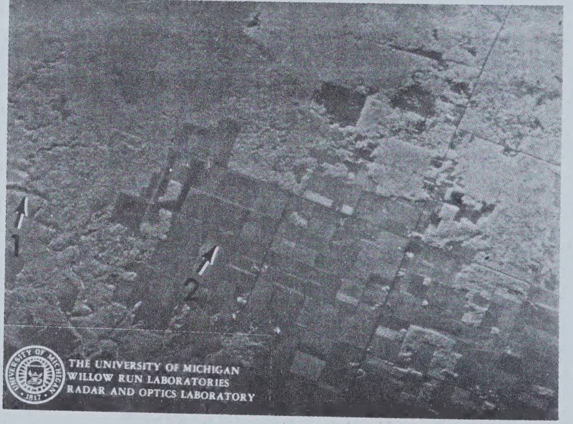

DIRECTION OF FLIGHT
N
1
2
3
5
4
THE UNIVERSITY OF MICHIGAN
WILLOW RUN LABORATORIES
RADAR AND OPTICS LABORATORY

B) WHITEFISH POINT MICHIGAN
1) Ice Structure Lake Superior
2) Radio Beacon
3) Whitefish Point
4) Breakwater
5) Prehistoric Coast Lines

FIGURE 4. SIDE-LOOKING RADAR IMAGERY AT 3 cm


1
2
3
DIRECTION OF FLIGHT

C) SAULT SAINT MARIE AIRPORT CANADA
1) Overhead Street Lighting Fixtures
2) Terminal Building and Parking Areas
3) Instrument Approach Lights


DIRECTION OF FLIGHT
1
2
THE UNIVERSITY OF MICHIGAN
WILLOW RUN LABORATORIES
RADAR AND OPTICS LABORATORY

D) TYPICAL RURAL AREA NORTHERN MICHIGAN
1) Drainage Patterns
2) Agricultural Areas

FIGURE 4. SIDE-LOOKING RADAR IMAGERY AT 3 cm

SUMMARY

During the past decade, the engineering design, construction, and operation of side-looking radar system has been perfected. These systems should now be given careful consideration and study in order to optimize their capability for various applications to environmental remote sensing. The hologram radar, on the other hand, is still in need of fundamental engineering development. These systems are usually burdened by their requirements for very long phased arrays and sophisticated motion-compensation systems. However, their downward-looking and range-contouring capabilities should be very useful, especially if they are used in conjunction with infrared or optical imaging systems.

Although they seem obvious, many of the more promising applications of microwave imaging to remote sensing are worth mentioning.

1) The mapping of remote, cloud-covered areas is straightforward.

2) The detection and composition of major geological features can indicate new areas of mineral exploration.

3) The detection of major ice cracks in large bodies of fresh water, such as the Great Lakes, offers the possibility of longer shipping seasons.

4) Very accurate profiles of the earth, moon, and planets can be obtained.

5) Some forms of pollution, for example oil slicks, may be detectable.

6) Agricultural surveys can be conducted to help determine the expected yield and status of crops on a national and possibly a global scale.

7) Measurements can be conducted to help determine the sea state.

8) Ship traffic can be accurately monitored on a local or global scale.

9) Imaging radar can serve as a tool in urban planning and development.

10) HF and VHF systems operating in lunar orbit enables us to detect subsurface features.

If microwave imaging devices are to be applied successfully to environmental remote sensing, two major tasks remain. First, the user must determine what radar parameters will optimize his objectives. Second, the developer must find more efficient and meaningful techniques for manipulating and processing multichannel radar data.

ACKNOWLEDGMENTS

Much of the side-looking radar work reported here has been sponsored by the National Aeronautics and Space Administration under Contract NAS-9-11036. The hologram radar work was accomplished under The University of Michigan's aerospace program, which is supported by the Air Force Avionics Laboratory, Wright-Patterson Air Force Base under Contract F33615-70-C-1433. The author personally wishes to thank Mr. Rendleman and Mr. Larson for their suggestions and encouragement.

REFERENCES

1. W. M. Brown, "Synthetic Aperture Radar," IEEE Trans. on Aerospace and Electronic Systems, AES-3, pp. 217-229, March 1967.
2. W. M. Brown and L. J. Porcello, "An Introduction to Synthetic-Aperture Radar," IEEE Spectrum, Vol. 6, No. 9, pp. 52-62, September 1969.
3. R. W. Larson, J. S. Zelenka, and E. L. Johansen, "Results Obtained from The University of Michigan Microwave Hologram Radar," Proceedings of the Seventh International Symposium on Remote Sensing of Environment, 17-21 May 1971.
4. R. W. Larson, J. S. Zelenka, and E. L. Johansen, "Microwave Hologram Radar," accepted for publication in IEEE Trans. on Aerospace and Electronic Systems, in 1972.
5. E. N. Leith, "Quasi-Holographic Techniques in the Microwave Region," Proceedings of the IEEE, Vol. 59, No. 9, pp. 1305-1318, September 1971.
6. W. M. Brown, Analysis of Linear Time-Invariant Systems, McGraw-Hill, 1963.
7. E. N. Leith and J. Upatnieks, "Reconstructed Wavefronts and Communication Theory," J. Opt. Soc. Am., 52, 1123 (1962).
8. E. N. Leith and J. Upatnieks, "Wavefront Reconstruction with Continuous-Tone Objects," J. Opt. Soc. Am., 53, 1377 (1963).
9. J. W. Goodman, Introduction to Fourier Optics, McGraw-Hill, 1968.

TOOLS FOR REMOTE SENSING EXPERIMENTS

GEORGE NAGY
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

The requirements for digital processing of Earth Resources Technology Satellite and Skylab imagery are described in terms of the experience gained with Gemini and Apollo photographs, televised weather satellite pictures, and airborne multispectral recordings. The topics discussed include format and statistical peculiarities of these data sets, analog-to-digital conversion and playback devices, image registration, linear and projective transformations, distortion correction, cloud detection, entry of ground-truth information, and the evaluation of classification results.

INTRODUCTION

This paper is intended as a chronicle of IBM's experience (primarily at the T. J. Watson Research Center) in the development of automatic digital image processing methods for remote sensing applications. This effort cannot be described as a coherent project leading to economically viable solutions to any given problem area. Nevertheless, it is believed that a sufficient number of important issues have been considered and experimentally investigated to warrant the preparation of a summary including a comprehensive guide to results published elsewhere.

While the work may be considered primarily as preparation for the kind of imagery expected from ERTS (Earth Resources Technology Satellites), SKYLAB (the manned satellite observatory) and high-flying NASA survey aircraft, several results of intrinsic interest, with possible application outside the field of remote sensing, have emerged in the course of the investigation.

The initial task was the collection of sufficient imagery in digital form. Since none of the currently available sources of data approximate closely the specified characteristics of the satellite RBV (return-beam vidicon) and MSS (multispectral scanner) systems, but each reflects several characteristics of interest, as many as possible of these sources were tapped.

The next step was the development of an adequate facility for the performance of the experimental work. Since once a few pictures have been digitized, a considerable amount of experimentation can be carried out before digitization of massive amounts of imagery becomes necessary, relatively little emphasis was placed on scanning hardware. Adequate display and hardcopy equipment, however, was deemed essential for observation of the results of processing experiments.

Whatever the prospective application, one of the most basic requirements is the determination of the exact earth-location of each segment of imagery. In principle, this may be achieved under certain conditions by cross-correlating key image elements with already located segments. Even with fast Fourier transform techniques it soon became clear that computationally this is not a practicable approach. It was shown, however, that the application of certain sequential similarity detection algorithms reduces the computational load by several orders of magnitude.

A matter related to registration is the measurement of parallax due to variations in the elevation of the terrain. An algorithm was developed to measure tree-height from low-altitude stereo-photography.

Once the location of the image has been determined, it is desirable to transform the picture to a standard set of map-coordinates, eliminating in the process every possible source of radiometric and geometric distortion.

The algorithm most suitable to perform the transformation must be chosen in function of the exact nature of the transformation; several such algorithms were developed.

In view of the inordinate data-collection capabilities of the proposed satellite system - it is estimated that each day's ERTS coverage would fill 250 reels of high density magnetic tape - image compaction procedures are of more than academic interest. Some results in this area were obtained, but it is clear that considerable improvement is necessary to make a significant dent in the data storage problem.

Automatic classification methods are essential for a large number of applications connected with resource surveys such as the timber census, crop prediction, and snow and ice studies. The approach taken emphasized the importance of the collection of ground-truth; the major objective was the reduction of the size of the required control sample through judicious use of the remotely sensed data.

Clouds may be treated either as obstructions to be automatically detected in order to avoid further processing of blighted pictures, or as objects of interest whose motion reveals the wind-velocity fields so important for accurate weather forecasting. The detection problem was tackled through the use of the sharp boundaries between brightly lit clouds and their shadows, and cross-correlation type algorithms were developed to measure cloud velocity and direction.

Each of these items will be discussed in additional detail in the following sections, but for a thorough treatment the reader is urged to turn to the references provided in the bibliography.

DATA

The four major sources of data utilized in our studies were weather satellite pictures, Apollo and Gemini photographs, multispectral observations from the University of Michigan airborne scanner, and low-altitude stereo color negatives.

1. United States Weather Satellite Vidicon Imagery [1]

Although at first both raw video and assembled mosaics from the ESSA-9 photographs were used, they were completely supplanted by the superior ITOS-1 (the current satellite) imagery obtained through the courtesy of the National Environmental Satellite Center.

Individual frames of uncorrected video data from the ITOS-1 satellite contain approximately 800 x 800 points at a resolution varying from 1.5 nautical miles at the subsatellite point to 3.0 n.m. at the edges of the frame. Each frame covers about 1700 x 1700 n.m. at the equator. The overlap between successive frames is about 50% and between adjacent orbits about 30% at the equator.

The video amplitude values are digitized to 256 levels, with 255 the brightest and 0 the darkest. Fiducial marks on the camera face plate, intended to allow removal of geometric camera distortion, appear in black in the video. Each frame contains nine fiducial marks.

The Advanced Vidicon Camera System (AVCS) data is recorded on 7 track, 556 bpi tape, with one orbit, containing 10 or 11 frames, to a tape. The block size is 5064 6-bit bytes, including 24 bytes of identification, and there are 800 East-West lines (4 to a record) in each frame (National Environmental Satellite Center Programmer note #140). Each data file is preceded by a header record containing the orbit number, and at the end of each tape there is ephemeris and timing information. Due to uncertainties in the A-D conversion, the number of elements per scan varies from scan to scan, up to a maximum of 945.

The following ITOS-1 orbits were used:

Orbit 2482 August 9, 1970
2494 August 10, 1970
2507 August 11, 1970
2519 August 12, 1970

These tapes cover North America, Mexico, and South America, at an average longitude of 100° West. For our purposes the most useful frames were the ones containing the relatively cloud-free region of Baja California; portions of the highly indented shoreline and several islands are clearly visible in each orbit.

2. Apollo and Gemini Photographs [2, 3, 4]

Most of the 2000 photographs collected on the six Gemini missions between 1964 and 1968 were obtained with hand-held cameras [2]. The astronauts appear to have favored high-oblique shots, with considerable variation in the scale and orientation of photographs of the same area. For registration experiments three frames containing Cape Kennedy (frames 5753, 45599, and 233544) were scanned. Two of these photographs are almost at the same scale and show little foreshortening at the coastline.

In addition to hand-held cameras, some of the Apollo missions were equipped with a bank of four boresighted 70mm Hasselbad cameras [3]. The SO-65 project on the Apollo-9, in particular, was designed to assess the capabilities and limitations of multi-band photography in a variety of applications. The satellite photography was carefully coordinated with aerial photography from almost a dozen aircraft flying at altitudes ranging from 3000 feet to 60,000 feet, with airborne multispectral scanner coverage, and with the simultaneous collection of ground-truth from several test sites [4].

For agricultural purposes the most popular test site appears to be the Imperial Valley of California. Extensive SO-65 coverage is available for this area, with overlapping frames both in the same orbit of the Apollo-9 and in successive orbits several days apart. Each photo-quadruplet consists of three black-and-white negatives in the green, red, and infrared, and of a false-color composite including all three of these bands. The scale of the vertical photography is about 1:1.5M. Only the red, green, and infrared negatives in frame A59-3700 (Salton Sea) were scanned and used. There is an offset of about 1/6 mile, due to boresight error between the green and the infrared pictures, but no significant rotational discrepancy.

3. Michigan Multispectral Observations [5,6]

The multispectral scanner is an electro-optical device for imaging radiation in a fixed set of spectral windows from an array of radiating (or reflecting) elements on the ground. The Michigan instrument is operated on an unpressurized C-47 airplane flying a straight level course.

The transverse component of the scan pattern is generated by a rotating axe-blade mirror assembly with a field of view limited to 40° on either side of nadir. The longitudinal component is obtained directly from the forward motion of the aircraft.

The detector assembly consists of a photomultiplier spectrometer with ten channels between 0.4μ m and 1.0μ m. The response of the photomultipliers for each channel is recorded on fourteen track 1" tape at 60 inches per second.

Three different calibration sources are viewed by the spectrometer during the 280° dead period of the scan, including a quartz-iodine lamp filtered to match the solar spectrum, an opal-glass plate at the top of the aircraft which integrates the diffused light, and a black reference which is just the dark interior of the scanner. The calibration signals are recorded on FM tape along with time marks and synchronization signals generated by a gyro-stabilized roll indicator.

After the flight the analog signal containing the ten-channel measurements is corrected for roll angle, digitized, unskewed, and normalized with respect to the calibration signals.

The data kindly provided for us by the Infrared and Optics Laboratory of the Institute of Science and Technology of the University of Michigan was collected under NASA auspices in conjunction with the Imperial Valley experiments mentioned above.

The two flight-lines selected were flown over McCabe Road, near the town of El Centro, California at altitudes of 5000 and 10,000 feet. The area covered by the 5000-foot flight is approximately 4000 acres, and by the 10,000-foot flight, 10,000 acres. After deleting a fraction of the scan-lines in order to retain an aspect ratio of 1:1, the former contained 345 scan-lines of 200 pixels each, and the latter, 200 scan-lines of 200 pixels. Since each pixel is a ten-dimensional vector, and eight bits are necessary for each component, a total of about 10^7 bits were retained for processing.

The major sources of distortion are uncorrected position and attitude errors, solar and atmospheric effects, and processing noise in the amplifying, recording, and analog-to-digital conversion equipment. Optical distortions introduced in the scanner itself are quite minor by comparison. Some statistical features of the data are described in reference [7].

4. Low-altitude Stereo-photographs [15]

These photographs of a wooded area were made on 70mm ANSCO/200 reversal film, but were converted prior to scanning to 35mm color separation negatives. The scale of the originals was 1:1580, and they were digitized at a resolution of approximately 2.7 ground feet per picture element. The film was provided to us by the United States Forest Service.

EXPERIMENTAL FACILITY [8,9,10,23]

The experimental image processing facility which has evolved in the course of the work is built around a console-oriented computer complex including a dedicated IBM 1800 control processor with high speed access to a Model 360/67 under TSS (Time Sharing System) and a Model 360/91 under OS/MVT (Operating System - Multiprogrammed with Variable Number of Tasks).

The features of primary interest to remote sensing experiments are two flying-spot scanners including the 35mm transparency scanner used to digitize the stereo-photographs, a cathode-ray tube Polaroid and 35mm film-output, a Stromberg-Datagraphix Charactron photoprinter with 1024 x 1024 addressable positions on 8" x 8" paper, a 32K core buffer and a 1M digital disk buffer serving either one color display or three high-resolution black-and-white displays capable of 484 x 448 element definition at 64 levels of grey and 968 x 864 elements in the binary mode, and a graphic tablet used for entering ground-truth.

This configuration is favourable to time-shared operation and has contributed to the development of three separate higher-level procedure-oriented picture processing languages. One system, based on 256 x 256 segments of imagery, supports itself on the elaborate filing organization of TSS. Another is primarily a high-speed processor for very large pictures (up to 4000 x 4000 elements) on the Model 91 [23].

The third is intended to expedite traffic between the Model 91 and the 1800, with commands and data on I/O operations issued at the 1800 end but the more time consuming operations performed on the large computer. Some use was also made of a fourth system, one designed for interactive processing of line drawings, for entering field boundaries in the multispectral data.

The various programming systems offer the usual image processing menu including enhancement, frequency domain filtering, translation, rotation, magnification, noise removal, gamma-correction, mosaicking, contouring, grey-slicing and density histograms.

In addition to the scanners available on the premises we have had access to large high-precision drum scanner/recorders at other IBM laboratories in Kingston, N. Y. and in Houston, Texas. These machines were used to digitize the Apollo and Gemini photographs at a resolution corresponding to a total of 4000 x 4000 elements over a 7" x 7" positive print, at 64 levels of grey-scale. At this resolution the printed out version of the digitized imagery is barely distinguishable from the original, and it is in fact questionable whether such a high sampling rate is justified by the information content of the photographs.

A false-color display system at the IBM Scientific Center in Houston has also been used to study the effectiveness of such a display in photointerpretation [11].

DETERMINATION OF ANCHOR POINTS [12,14]

The usual method for registering the corresponding segments of two pictures of the same area is to calculate the correlation coefficient between a window taken from one picture and an area of the same size taken from the other picture. The largest value of this coefficient over the search space is accepted as corresponding to the desired coordinate location [12].

When the search space is large, due to large initial uncertainty in the location of the window or to the presence of rotational and scale variations in addition to simple translation, the amount of computation required by this approach is enormous. While considerable reduction in the number of operations performed may be achieved, especially for window areas containing a large number of elements, through the use of the fast Fourier transform, the operation is still too slow for application to ERTS data through general-purpose digital computers [13].

It can be shown, however, that only in a relatively small number of positions is it necessary to compute the contribution of many elements in the window to the total figure of merit. In most instances, consideration of only a few points reveals that the displacement under consideration will not yield an acceptable match between the two pictures, and the algorithm may then proceed to the next trial position.

Formalization of this idea yields the class of sequential similarity detection algorithms [14], which were studied from a theoretical viewpoint under the assumption that the deviation between the two pictures (or noise) is independently and exponentially distributed. The experimental work, which shows rather close agreement with the theory, was carried out on the weather satellite pictures described earlier. The number of operations required by different search spaces and window areas was tabulated; by way of example, for a search area of dimensions 2048 x 2048 and a window size of 256 x 256 elements, the direct method requires 10^{12} operations, the FFT 8.8×10^9 operations, and the best of the sequential algorithms, 2.3×10^7 operations.

PARALLAX MEASUREMENTS [15]

The determination of parallax from stereo-photographs differs from the registration problem just discussed in that the displacement function may vary quite rapidly from point to point, therefore the use of window functions based on uniform displacements is inappropriate.

Good results in estimating parallax on a single stereo-pair were obtained with the assumptions that the difference in the grey-levels of corresponding trial-pairs of points in the two pictures is a zero-mean gaussian process and that the departure of the displacement-function from its average value over a small neighbourhood is also of the same form. Since in this formulation the grey-level difference distribution is conditional upon the parallax, for which the gaussian continuity condition gives an *a priori* distribution, Bayes rule may be invoked to estimate the *a posteriori* dependence of the parallax on the observed video distribution.

The computation of the altitude contour lines from the parallax information is a straightforward problem in analytic geometry. Some results of applying the complete procedure to the estimation of forest-tree heights are shown in reference [15].

DISTORTION CORRECTION [12,16,17,18,19,20,21,29]

Most of the work in distortion correction was based on the types of distortion expected to be encountered in the ERTS imagery. Both geometric distortion, resulting from sensor misalignment, deviations introduced within the sensor itself, and analog recording and transmission errors, and radiometric distortion, due to external variation in the illumination, to motion smear, and to modulation-transfer-function defects, were considered. A comprehensive analysis of the expected sources of error is included in [18].

Experimental work was concentrated on the correction of "small" geometric deviations corrected on the basis of the automatically extracted location of the nine reseau marks etched on the face of the three return-beam vidicon tubes [19]. The imagery was simulated by means of Apollo-9 photographs quantized at 4096 x 4096 picture elements to six levels of grey.

The "point shift" algorithm developed for this purpose effects considerable saving in computer time over direct implementation of the required transformation provided the major underlying assumption, that the correction function is relatively smooth, is satisfied. If this is indeed the case, then a number of adjacent (in both dimensions) output points may be assigned the video-values of a series of adjacent input points. The maximum positional error is held to ± 1/2 picture element. Such a transformation preserves the spatial frequency characteristics of the original image, and can be performed rapidly on byte-string oriented computers because the inner loop computation is replaced by a simple MOVE instruction [17,18,19].

For transformations requiring a change of scale or other situations where the above assumption is not satisfied, high-speed methods of performing linear and perspective transformations were also developed, but not tested to any significant extent [20]. Some properties of bicubic splines, for image interpolation, are also reported in [2].

Estimates of the amount of computation required for various degrees of geometric and photometric correction are tabulated in [16].

IMAGE COMPACTION [22]

In view of the National Oceanographic and Atmospheric Agency's interest in storing large quantities of "high-resolution" vidicon data, both in raw and in processed form, an attempt was made to adapt video compaction techniques to the ESSA-9 and ITOS-1 imagery.

On grounds that Fourier filtering in various guises would be an integral part of most processing schemes, theoretical work centered on the selection and quantization of the most informative Fourier coefficients of a transformed image. An optimal rule was derived for minimizing a generalized linear error function comprising a weighted sum of the video components and their spatial integrals and derivatives.

Experimentally, visually acceptable results were obtained at a compaction rate corresponding to storing one bit for each original picture element. This is comparable to results obtained both within IBM and elsewhere with a variety of compaction algorithms, but in view of the computational load of even the Fast Fourier Transform compared to "local operators" the method seems warranted only if the Fourier coefficients have to be computed for some other purpose in any case.

CROP CLASSIFICATION [7,24]

The point of departure of our work from other methods of automatic crop classification by means of multispectral signatures is the concept of utilizing the information collected by airborne sensors to determine the areas for the collection of ground truth. By this expedient we avoid the possibility of basing the classification on irrelevent or outdated information and at the same time minimize the laborious process of labelling training areas with the appropriate crop identity tags by means of visual examination or other ancillary data.

The method for the determination of homogeneous areas (which are expected to contain either a single crop or species not readily distinguishable at the spectral and spatial resolution of the sensor) is a variant of a well-known clustering algorithm. In order to improve processing speed - a paramount consideration here as in most other remote sensing schemes - the algorithm was modified to take advantage of certain continuity conditions based on agricultural planting practices. Once the homogeneous (but not necessarily contiguous) regions have been delineated by the computer, only a small fraction of each type of region need be sampled by a ground-crew or otherwise identified. The combination of the cluster configuration and the crop-category information is a straightforward and rapid table look-up operation.

This approach was tested on the seventeen crop and terrain categories contained in the Michigan data described earlier. Although direct comparison with the standard supervised-learning method is difficult because of the necessary choice of training sets, our method did prove far superior to a sampling scheme using randomly chosen test plots of the same average size as the machine-determined clusters.

As expected, further improvement can be obtained by reducing the number of classes and by performing the classification on a "per-field" basis. We were not, however, successful in developing a satisfactory method for deriving the complete field boundaries automatically.

CLOUDS [25,26,27]

The estimation of prevailing wind direction from satellite cloud pictures is the unremitting daily task of meteorologists preparing input data for the currently operational national weather forecasting program.

Attempts to automate this process, in progress as a joint NOAA - IBM project since 1963, are beginning to produce useful results at a throughput rate approaching operational standards.

The basic process consists of cross-correlating the low-resolution imagery derived at several hours' intervals from the geosynchronous high-altitude (20,000 nautical miles) satellites generally stationed over the Atlantic and Pacific Oceans. Detailed implementation and interpretation of the results is complicated by the sometimes rapid deformation of cloud masses, by the shadows cast by the topmost layers, and by the variation in wind velocity with altitude.

The sharp delineation of brightly lit clouds against their dark shadows on the ground or on other clouds has been exploited for detecting the presence of such clouds on the much higher resolution Apollo photographs. Although the direction of the shadows depends only on the sun-angle, their distance from the cloud is a function of the cloud altitude as well. The method is applicable only to intermittent cloud-cover and yields but a modest improvement over results obtained by relying on the brightness as a detection criterion.

CONCLUSION

Of the over one hundred potential application areas for satellite data which have been identified, none has emerged so far where satellite surveillance has been conclusively demonstrated to have an economic edge over alternative methods. It is only through the combined benefits accruing from many projects that the current undertaking can possibly be justified.

With few exceptions, the work described in this paper reflects the magnitude and diversity of the problems for which successful solutions must be developed even before meaningful experimentation on the automatic extraction of information from satellite imagery for specific applications can begin in earnest. While automatic processing is the ultimate goal, it is quite likely that useful results will continue to be obtained with standard visual photointerpretation techniques for many years before automatic techniques become truly competitive.

Realistic experimentation on individual subject areas requires a very large investment in terms of equipment and programming; it would seem that extensive cooperation between the various institutions engaged in this work should be well worth the extra effort. Yet scanning the literature reveals very little in the way of cooperative behaviour, except for the exchange of data, among different institutions; instead, each group or department seems bent on developing a complete system, even though some of the components may be used only once or twice and then only tangentially with respect to the major direction of the research effort. It must be admitted, however, that we have not been able to achieve a completely satisfactory degree of standardization with respect to data formats, programming systems, and hardware interfaces even within our own organization, and considerable energy is still devoted to the duplication of conceptually trivial tasks.

We believe that further progress is conditional upon the extent to which the fragmented experimentation that is the order of the day is replaced by a cohesive effort leading to efficient utilization of the massive resources available. Whether such conscientious standardization and voluntary sharing of the myriad essential data manipulation routines is possible without a strong external leadership remains to be seen.

REFERENCES

1. J. A. Leese and A. L. Booth, "Archiving and Climatological Application of Meteorological Satellite Data", Report of the U.S. Dept. of Commerce, Environmental Services Administration, Rockville, Md., prepared for the Fifth Session of the WMO Commission for Climatology, Geneva, Oct. 1969.

2. National Aeronautics and Space Administration, Earth Photographs from Gemini VI Through XII, NASA SP-171, Washington, 1969.

3. NASA: Apollo-9 Photographic Plotting and Indexing Report, prepared by Mapping Science Laboratory for Earth Resources Division, Science and Applications Directorate, Manned Spacecraft Center, Houston, October 1969.

4. NASA-MSC Earth Resources Aircraft Program, Mission 89 Summary Report, Houston, Texas, June 1969.

5. R. F. Nalepka, "Investigation of Multispectral Discrimination Techniques", Optics Laboratory of the Institute of Science and Technology of the University of Michigan Report 2264-F, Ann Arbor, Michigan, 1970.

6. R. E. Marshall and F. J. Kriegler, "An Operational Multispectral Survey System", Seventh Michigan Symposium on Remote Sensing, Ann Arbor, Michigan, May 1971.

7. G. Nagy, G. Shelton and J. Tolaba, "Procedural Questions in Signature Analysis", Seventh Michigan Symposium on Remote Sensing, Ann Arbor, Michigan May 1971; also IBM Report RC 3517, Yorktown Heights, N.Y., Sept. 1971.

8. N. M. Herbst and P. M. Will, "An Experimental Laboratory for Pattern Recognition and Signal Processing", accepted for publication in Comm. ACM.

9. N. M. Herbst and P. M. Will, "Design of an Experimental Laboratory for Signal and Image Processing", International Symposium on Computer Graphics, Brunel, U. K., April 14, 1970, in Advanced Computer Graphics, Parslow & Green (Ed.), Plenum Press, New York, 1971.

10. N. M. Herbst and P. M. Will, "Design of an Experimental Laboratory for Pattern Recognition and Signal Processing", IBM Report RC 2619, Yorktown Heights, N. Y., September, 1969.

11. P. R. Tobias and R. E. Jensen, "Pseudocolor Enhancement of Image Perception", IBM Report 320-2386, Houston Scientific Center, April 1970.

12. R. Bakis, M. Wesley and P. Will, "Digital Correction of Geometric and Radiometric Errors in ERTS Data", Seventh Michigan Symposium on Remote Sensing, Ann Arbor, Michigan, May 1971.

13. H. F. Silverman, "On the Use of Transforms for Satellite Image Processing", Seventh International Symposium on Remote Sensing of Environment, May 17-21, Ann Arbor, Michigan, also IBM Report RC3348 Yorktown Heights, N.Y., May 1971.

14. D. Barnea and H. F. Silverman, "A Class of Sequential Similiarity Detection Algorithms for Fast Digital Image Registration", accepted for publication in IEEE-C, April 1972; also IBM Report RC3358, Yorktown Heights, N. Y., May 1971.

15. R. Bakis and P. Langley, "A Method for Determining the Height of Forest Trees Automatically from Digitized 70mm Color Aerial Photographs", Seventh

International Symposium on Remote Sensing of Environment, Ann Arbor, Michigan May 1971; also IBM Report RC 3387, Yorktown Heights, N.Y., April 1971.

16. R. Bakis, M. Wesley, P. Will, "Digital Image Processing for the Earth Resources Technology Satellite Data", presented at American Society of Photogrammetry & American Congress on Surveying and Mapping Convention, Washington, March 7-12, 1971.

17. R. Bakis, M. Wesley, P. Will, "Digital Image Processing for the Earth Resources Technology Satellite Data", IFIP Congress 1971, Ljubljana, Yugoslavia, August 1971.

18. P. M. Will, R. Bakis, M. A. Wesley, "On an All-Digital Approach to Earth Resources Satellite Image Processing", IBM Report RC 3027, Yorktown Heights, N.Y., September 1970.

19. H. Markarian, R. Bernstein, D. G. Ferneyhough, L. Gregg, F. S. Sharp, "Implementation of Digital Techniques for Correcting High Resolution Images", ASP - ACSM Fall Convention, San Francisco, Calif., September 1971; also IBM Report FSC 71-0012, Gaithersburg, Md., July 1971.

20. M. Szeto, "An Approximation for Projective Transformations for Satellite Imagery", IBM Report RC 3320, Yorktown Heights, N.Y., April 1971.

21. C. Lindstedt, "Preliminary Experiments on Geometric Transformations of Images", IBM Report RC 3314, Yorktown Heights, N.Y., April 1971.

22. H. F. Silverman, "Optimal Fourier Domain Data Compaction for Earth Resources Satellite Data" presented at the Seventh International Symposium on Automatic Photointerpretation & Recognition, December 1-2, 1970, Baltimore, Md.; also "An Optimizing Fourier Domain Compaction Method for ERTS Data", IBM Report RC 3277, Yorktown Heights, N.Y., March 1971.

23. P. M. Will, G. M. Koppelman, "MFIPS: A Multifunctional Digital Image Processing System", IBM Report RC 3313, Yorktown Heights, N.Y., April 1971.

24. G. Nagy and J. Tolaba, "Nonsupervised Crop Classification Through Airborne Multispectral Observations", IBM Report in press, Yorktown Heights, N.Y.

25. J. A. Leese, C. S. Novak, and V. R. Taylor, "The Determination of Cloud Pattern Motion from Geosynchronous Satellite Image Data", Pattern Recognition, Vol. 2, #4, December 1970.

26. J. A. Leese, C. S. Novak, and V. R. Taylor, "The Determination of Cloud Pattern Motion from Geosynchronous Satellite Image Data", ASP Symposium on Computational Photogrammetry, Falls Church, Va., January 1970.

27. J. A. Leese and B. Clark, "An Automated Technique for Obtaining Cloud Motion from Geosynchronous Satellite Data Using Cross Correlation", J. Applied Meteorology, Vol. 10, #1, pp. 110-132, February 1971.

28. L. M. Mumbower, "Automated Processing for Earth Resources", Proceedings of the Princeton University Conference on Aerospace Methods for Revealing and Evaluating Earth Resources, Princeton, N.J., June 1970.

29. R. Bernstein, H. Branning, D. Ferneyhough, "Geometric and Radiometric Correction of High Resolution Images by Digital Image Processing Techniques", IEEE Annual International Geoscience Electronics Symposium, Washington, D.C. August 1971.

SOME ANALYTICAL AND NUMERICAL TECHNIQUES FOR REMOTE SENSING

R. MITTRA, D. H. SCHAUBERT AND M. MOSTAFAVI
University of Illinois
Department of Electrical Engineering
Urbana, Illinois 61801

ABSTRACT

The remote sensing problem is formulated in terms of the operator equation Lu = 0 where L is a linear differential operator with unknown coefficients. The problem consists of determining these unknown coefficient functions when the boundary value of u is known as a function of spatial frequency. Three methods for doing this are discussed and some examples are presented.

I. INTRODUCTION

Much of the information we have about our world is obtained via remote sensing measurements. We use remote sensing techniques whenever it is undersirable, impractical or impossible to place measuring probes directly at the point being investigated. This includes microwave reflectivity measurements to determine areas where crops are diseased[1], radio occultation measurements for studying the atmospheres of other planets[2], radio soundings to investigate the ionosphere[3], particle scattering measurements to determine quantum mechanical potential functions[4], seismic measurements to locate mineral deposits[5], ultrasonic probing of the human body to diagnose diseased and damaged tissues[6], and many more.

A common trait of all remote sensing experiments is that the data obtained by measurements must be processed to obtain the information we desire. The type of processing which is required depends upon the particular experiment. The particular type of problem which will be discussed in this paper is that of determining the electrical permittivity of a medium which is uniform in the x and y directions and varies according to the profile function, $\kappa(z)$, in the z direction.

Using the techniques of Gelfand and Levitan[7] and of Marchenko[8], many authors, e.g., Becker and Sharpe[9], have reported the solution of this type of problem. All of their work, however, requires that the response of the medium be known as a function of the temporal frequency. It is also required that the medium itself be independent of frequency, i.e., nondispersive. Often, however, it is desirable to make measurements at a single frequency. Therefore, in this paper we consider the remote probing problem in the angular spectrum domain. Here the independent variable, β, is the spatial frequency which can be associated with $k \sin \theta$ where k is the free space wave number and θ is the angle of incidence of a probing plane wave.

The technique by which the response of the medium is measured is to illuminate the medium using a spatially confined source. The total field at the interface is then measured. The information is then converted to the angular spectrum domain by taking the Fourier transform with respect to the x direction.

In the next section, the differential equation describing the problem is derived. Three different methods of obtaining the profile are then discussed. The first method makes use of the form of the equation in the limit of low temporal frequencies[10]. The second method can be used for arbitrary frequency and is based on a new representation for the wave function which is appropriate for the angular spectrum domain. The third method is a direct numerical approach which uses parameter optimization techniques[11].

II. FORMULATION OF THE PROBLEM

The geometry of the problem to be considered is shown in Figure 1. Note that there are no variations with respect to the y direction. The dielectric is also uniform in the x direction and varies in the z direction as $\kappa(z)\varepsilon_o$ where ε_o is the permittivity of free space. The permeability is assumed to be constant and equal to that of free space. It is assumed that a known electromagnetic wave of a single frequency is incident from the left on the air-dielectric interface at z = 0 and that the tangential fields at the surface of the dielectric can be measured for all values of x. Of course, the actual measurement could be performed at some plane $z = \alpha$, $\alpha < 0$, and then the fields at the interface could be computed since the propagation in free space is completely known.

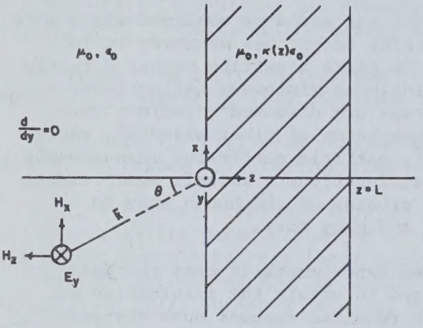


Figure 1. Remote probing problem in the angular spectrum domain.

If the incident field is polarized such that the components E_x, E_z, H_y are all zero, then propagation in the medium is completely described by the scalar equation

$$\nabla^2 v(x,z) + \omega^2 \mu_o \kappa(z) \varepsilon_o v(x,z) = 0 \tag{1}$$

(ω is the angular frequency of the incident wave). The function $v(x,z)$ may be identified with E_y or H_z, which ever is more convenient for the problem at hand.

In order to obtain the differential equation in the angular spectrum domain, we introduce the Fourier transform relations:

$$u(\beta,z) = \int_{-\infty}^{\infty} v(x,z) \exp(-j\beta x)\, dx \tag{2a}$$

$$v(x,z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} u(\beta,z) \exp(j\beta x)\, d\beta \tag{2b}$$

Then (1) becomes

$$\frac{d^2 u(\beta,z)}{dz^2} + [\omega^2 \mu_o \kappa(z) \varepsilon_o - \beta^2]\, u(\beta,z) = 0 \tag{3}$$

This is the differential equation we are seeking. Recall that ω, μ_o, and ε_o are constants, β is an independent variable, and $\kappa(z)$ is the unknown profile function.

If we define a linear second order differential operator

$$L = d^2/dz^2 + \omega^2\mu_o\kappa(z)\varepsilon_o - \beta^2 \tag{4}$$

we can write (3) as

$$Lu(\beta,z) = 0 \tag{5}$$

We can now see why the problem we are considering is referred to as an "inverse problem". The "forward problem" consists of determining $u(\beta,z)$ when L is given. The inverse problem consists of determining L when we know something about $u(\beta,z)$. For our particular problem, we must determine the unknown coefficient function $\kappa(z)$.

III. METHOD 1: A LINEAR INTEGRAL EQUATION

In this method, we assume that the medium is terminated at $z = L$ by a perfect electric conducting wall (see Figure 2). This simplifies the expressions by causing some of the boundary terms to vanish.

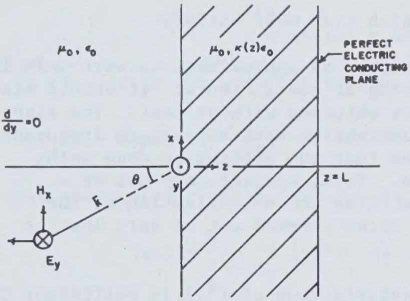


Figure 2. Geometry of the remote probing problem with a perfectly conducting plane at z = L.

We look now at (3) and consider the special case of $\omega = 0$.

$$\frac{d^2S(\beta,z)}{dz^2} - \beta^2 S(\beta,z) = 0 \tag{6}$$

Note that $S(\beta,z)$ is the static solution for $H_z(\beta,z)$ and is independent of $\kappa(z)$. Therefore, $S(\beta,z)$ can be calculated.

$$S(\beta,z) = S_o \sinh[\beta(z-L)] \tag{7}$$

We can now use (3), (6), and (7) and the fact that both $u(\beta,z)$ and $S(\beta,z)$ vanish at $z = L$, to arrive at the following integral equation for $\kappa(z)$.

$$\omega^2\mu_o\varepsilon_o \int_o^L \kappa(z)\, u(\beta,z) \sinh[\beta(z-L)]\, dz \tag{8}$$

$$= -\beta \cosh[\beta L]\, u(\beta,o) - \sinh[\beta L] \frac{du(\beta,o)}{dz}$$

Two points should be noted about (8). First, the right hand side of the equation is related to the tangential fields at the surface of the dielectric and, thus, can be determined from the input impedance. Second, Equation (8) is nonlinear because $u(\beta,z)$ depends on $\kappa(z)$. Therefore, there is no systematic procedure which can be used to find $\kappa(z)$ from (8). To circumvent this problem, a linear form of the integral equation is obtained by expanding $u(\beta,z)$ in a power series of ω, the angular frequency.

$$u(\beta,z) = \sum_{n=0}^{\infty} C_n(\beta,z)\omega^n, \quad C_n(\beta,z) = \frac{1}{n!} \left. \frac{\partial^n u(\beta,z)}{\partial\omega^n} \right|_{\omega = 0} \tag{9}$$

It can be proven that for ω sufficiently small, no surface wave of this polarization will propagate and that (9) is a uniformly convergent series. This allows us to substitute (9) into (8) and interchange the order of summation and integration. Equating the coefficients of like powers of ω yields a series of equations. The ω^2 term, in particular, yields

$$\mu_o\varepsilon_o \int_o^L \kappa(z)\, C_o(\beta,z) \sinh[\beta(z-L)]\, dz = -\beta \cosh[\beta L]\, C_2(\beta,0) - \sinh[\beta L] \left. \frac{dC_2(\beta,z)}{dz} \right|_{z = 0} \tag{10}$$

To simplify (10), note that $C_o(\beta,z) = u(\beta,z)\Big|_{\omega = 0}$. But, for $\omega = 0$, (3) becomes identical with (6) and $u(\beta,z)$ becomes identical with $S(\beta,z)$. Therefore, (10) can be written

$$\mu_o\varepsilon_o \int_o^L \kappa(z) \sinh^2\beta(z-L)\, dz = -\beta \cosh[\beta L]\, C_2(\beta,0) - \sinh[\beta L] \left. \frac{dC_2(\beta,z)}{dz} \right|_{z = 0} \tag{11}$$

This is the linear integral equation which $\kappa(z)$ must satisfy.

It should be noted that (11) is exact. No approximations were made in the expansion of $u(\beta,z)$ or in the matching of coefficients. It should also be noted that the linear equation is not obtained without cost. The right hand side of (11) involves the second derivative with respect to frequency of the tangential fields. This requires that the probing be done using frequencies in the neighborhood of zero. This, however, may not be a limitation since some remote probing, such as the determination of the electrical properties of the earth, is often carried out at very low frequencies.

It should also be noted that the satisfaction of (11) is sufficient to guarantee that the equations obtained by matching other powers of ω are all satisfied. Furthermore, it can be proven that (11) has a unique solution within the class of Laplace transformable functions.

A final comment about the numerical behavior of the solution to (11) is needed. The matrices obtained by various quadratures on (11) were found to be ill-conditioned. That is, small deviations in the data appearing on the right hand side of the equation lead to large deviations in the calculated profile function. This limits the application of this method in practical problems. However, the Equation (11) can be proven to be nonsingular, and hence, better inversion methods than were used in our investigation could lead to usable results.

IV. METHOD 2: A NEW REPRESENTATION

Recall that method 1 required the angular spectrum response of the medium to be obtained for angular frequencies in the vicinity of $\omega = 0$. The method presented in this section can be used for any fixed nonzero value of ω.

Before continuing, we introduce an equivalent form of (3) which will be used as the basic differential equation in this section

$$u''(\gamma,z) + \gamma^2 u(\gamma,z) = Q(z)\, u(\gamma,z) \tag{12}$$

where

$$\gamma^2 = k^2 - \beta^2$$

$$Q(z) = -k^2[\kappa(z)-1]$$

$$k^2 = \omega^2 \mu_o \varepsilon_o$$

The prime denotes differentiation with respect to z. Note that if we determine $Q(z)$, $\kappa(z)$ can easily be found.

We will assume that $Q(z)$ is bounded and that

$$\int_o^\infty z|Q(z)|\, dz < \infty \tag{13}$$

$$\lim_{z\to\infty} u(\gamma,z) = \exp[i\gamma z] \tag{14}$$

We then write the solution to (12) in the form

$$u(\gamma,z) = \exp[i\gamma z] + \frac{1}{2\pi i} \int_{-\infty}^{\infty} u(\eta,z)\, \exp[i(\gamma + \eta)z]\, (\gamma + \eta)^{-1} P(\eta)\, d\eta \tag{15}$$

The associated profile function is then given by the expression

$$Q(z) = \frac{1}{\pi i} \int_{-\infty}^{\infty} [iu'(\eta,z) - \eta\, u(\eta,z)]\, \exp[i\eta z] P(\eta)\, d\eta \tag{16}$$

The function $P(\eta)$ which appears in (15) and (16) contains information about the medium being studied. We can determine $P(\eta)$ from a knowledge of $u(\gamma,o)$ by considering (15) for $z = 0$. That is, the equation

$$u(\gamma,o) = 1 + \frac{1}{2\pi i} \int_{-\infty}^{\infty} u(\eta,o)\, (\gamma + \eta)^{-1} P(\eta)\, d\eta \tag{17}$$

is a linear integral equation for $P(\eta)$.

Using these equations, we obtain the unknown function $Q(z)$ by the following steps:

- i) knowing $u(\gamma,o)$, solve (17) for $P(\eta)$
- ii) for various fixed values of z, solve (15) for $u(\gamma,z)$
- iii) obtain $Q(z)$ from (16).

It can be shown that (15) is a solution to (12). This is done by substituting (15) into (12). This results in an integral equation of the form

$$T(\gamma,z) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} T(\eta,z) \exp[i(\gamma + \eta)z] \, (\gamma + \eta)^{-1} P(\eta) \, d\eta \tag{18}$$

where $T(\gamma,z) = u''(\gamma,z) + \gamma^2 u(\gamma,z) - Q(z)\,u(\gamma,z)$. For functions of the type encountered in this problem, (18) has only the solution $T(\gamma,z) \equiv 0$. Therefore, (12) is satisfied.

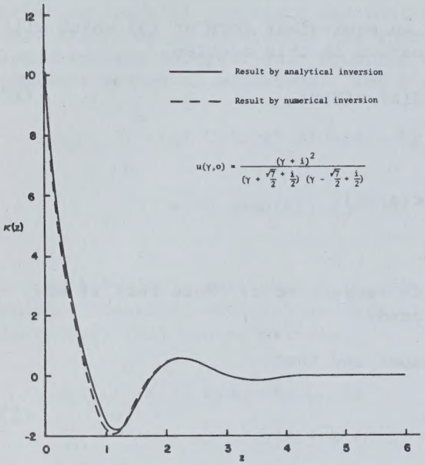


Figure 3. Results obtained when method 2 was applied to Sharpe's example.

Using an example due to Sharpe[12], the above method was tested for numerical stability. In this case, the field at the interface was given by

$$u(\gamma,o) = \frac{(\gamma + i)^2}{(\gamma + \frac{\sqrt{7}}{2} + \frac{i}{2})(\gamma - \frac{\sqrt{7}}{2} + \frac{i}{2})} \tag{19}$$

Figure 3 indicates the close comparison between the calculated result and the analytical result obtained by Sharpe.

V. METHOD 3: PARAMETER OPTIMIZATION

In contrast to the two methods presented above, the third method is numerical in nature. This method is more general than the previous ones, but it does lack the analytical niceties that are present in those methods. For instance, it is impossible to prove uniqueness of the solution. Nevertheless, this method can use data obtained in the angular spectrum domain or the frequency domain, or both, and it can be generalized to a wider class of geometries, e.g., circular columns and sheath configurations in the neighborhood of space vehicles. In addition, this method is numerically stable.

The method is based on the parameter optimization approach which is commonly used in systems theory. The method seeks to minimize a performance index function F which is defined as the L_2 norm of the error of the response.

$$F(\kappa_1, \kappa_2, \ldots, \kappa_m) = \sum_{i=1}^{P} \sum_{j=1}^{Q} \left| u^g(\omega_i, \beta_j, \kappa) - u(\omega_i, \beta_j, \kappa_1, \kappa_2, \ldots, \kappa_m) \right|^2 \tag{20}$$

where

κ_i are unknown parameters introduced to describe the medium,

$u^g(\omega_i, \beta_j, \kappa)$ = surface field for the true medium described by a profile function $\kappa(z)$ when measured for angular frequency ω_i and spatial frequency β_j.

$u(\omega_i, \beta_j, \kappa_1, \kappa_2, \ldots, \kappa_m)$ = surface field for a trial medium described by the parameters $\kappa_1, \kappa_2, \ldots, \kappa_m$ when measured for angular frequency ω_i and spatial frequency β_j.

The terms $u^g(\omega_i, \beta_j, \kappa)$ are the measured values of the field for the medium being investigated. In some cases the performance index also contains terms involving the errors in du/dz.

The strategy of the method is as follows. One begins with a continuous or layered model for the medium involving the parameters $\kappa_1, \kappa_2, \ldots, \kappa_m$. With the parameters set equal to some reasonable values, the response of this trial medium for each ω_i and β_j is calculated using (3) and some standard numerical technique for solving differential equations. This response is the $u(\omega_i, \beta_j, \kappa_1, \kappa_2, \ldots, \kappa_m)$ term in (20). Once the performance index F is calculated, the optimization algorithm determines a new set of parameter values and the process is repeated until a preset stopping criterion is satisfied, e.g. F is less than a given error tolerance ε.

The optimization technique has been applied to a variety of inversion problems involving inhomogeneous media with continuous as well as discrete variations in the profile function, for both planar and circular geometries.

The numerical calculations were done using two different optimization algorithms. Figure 4 shows the results obtained using Rosenbrock's method[13] for a profile function $\kappa(z) = \exp(2z)$. Figure 5 shows the results obtained using the conjugate gradients method[14] on the same profile. Figure 6 shows

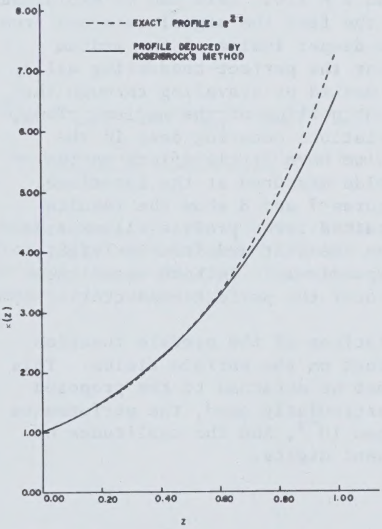


Figure 4. Profile of dielectric constant $\kappa(z) = e^{2z}$ for a slab terminated by a perfect conductor at $z = 1$. Rosenbrock's method.

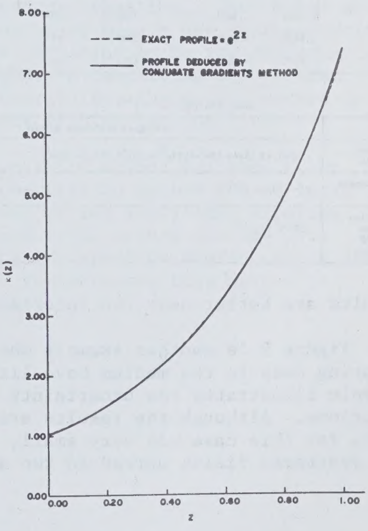


Figure 5. Profile of dielectric constant $\kappa(z) = e^{2z}$ for a slab terminated by a perfect conductor at $z = 1$. Conjugate gradients method.

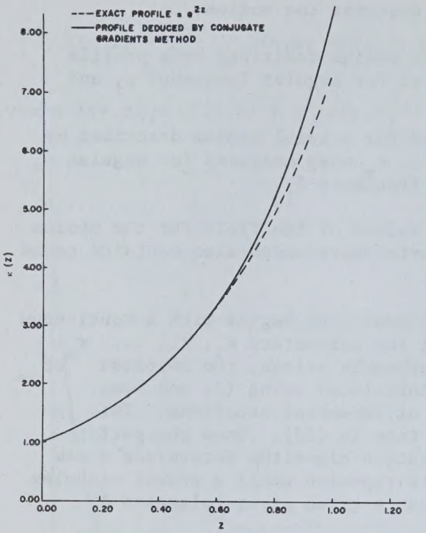


Figure 6. Profile of dielectric constant $\kappa(z) = e^{2z}$ for a slab terminated by a perfect conductor at $z = 1$. Random noise was added to the simulated data.

Table 1. Comparison of the Total Electric and Magnetic Field Amplitudes for the Exact and Inverted Profiles at $z = 0$. $k_0 = 1$.

Incident Angle in degrees	Amplitude of Total E Field			$\frac{\partial E}{\partial z} \alpha$ Amplitudes of the Total H Field		
	Exact	Rosenbrock's Method	Conjugate Gradients Method	Exact	Rosenbrock's Method	Conjugate Gradients Method
0	-.5692	-.5748	-.5693	.1308	.1293	.1313
30	-.5998	-.6057	-.5999	.2108	.2099	.2114
60	-.6637	-.6699	-.6638	.3831	.3835	.3837
90	-.6969	-.7034	- .6970	.4756	.4766	.4761

Exact profile $\kappa(z) = e^{2z}$

	Rosenbrock's Method	Conjugate Gradients Method
Inverted Profile	$\kappa(z)=1.0+2.16z+2.16z^2 +.58z^3+.726z^4$	$\kappa(z)=1.0+2.07z-1.64z^2 +1.42z^3+1.3z^4$
Performance Index	$F=.89 \times 10^{-5}$	$F=.136 \times 10^{-8}$
Execution time on FORTRAN Compiler	40.11 sec	35.15 sec.

Table 2. Comparison of the Total Electric and Magnetic Field Amplitudes at $z = 0$ for the Exact and Inverted Profiles Adding Random Noise to Simulated Data. $k_0 = 1$.

Incident Angle in degrees	Amplitude of Total E Field		$\frac{\partial E}{\partial z} \alpha$ Amplitude of Total H Field	
	Exact	Conjugate Gradients Method	Exact	Conjugate Gradients Method
0	-.5692	-.5584	.1307	.1209
30	-.5999	-.5887	.2108	.1998
60	-.6638	-.6518	.3832	.3696
90	-.6970	-.6847	.4757	.4607

Exact Profile $\kappa(z) = e^{2z}$

	Conjugate Gradients Method
Inverted Profile	$\kappa(z)=1.+2.11z+1.68z^2+1.25z^3+1.02z^4+.771z^5+.594z^6$
Performance Index	$F = .234 \times 10^{-1}$
Execution time on FORTRAN Compiler	133.71 sec.

the results obtained using the conjugate gradients method when the data was polluted by noise. The very good agreement here indicates that this method is stable enough to be used with laboratory data. Tables 1 and 2 contain information about these cases.

It can be seen that the computed results agree with the true profile better in the vicinity of $z = 0$ than $z = 1.0$. This can be explained by the fact the signal returned from the deeper inside of the medium (near the perfect conducting wall) is masked by traveling through the front portion of the medium. Thus, variations occuring deep in the medium have little effect on the fields measured at the interface. Figures 7 and 8 show the results obtained for a profile illuminated from the left and from the right, respectively. In both cases the results are better near the interface than near the perfect conductor.

Figure 9 is another example where variations of the profile function occuring deep in the medium have little effect on the surface fields. This example illustrates the uncertainty that must be attached to the proposed solutions. Although the results are not particularly good, the performance index for this case was very small, less than 10^{-3}, and the amplitudes of the scattered fields agreed to two significant digits.

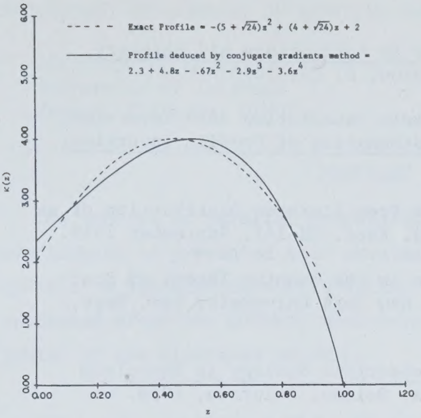


Figure 7. Profile of dielectric constant $\kappa(z) = -(5 + \sqrt{24})z^2 + (4 + \sqrt{24})\ z + 2$ for a slab terminated by a perfect conductor at $z = 1$. Plane wave was incident from $-z$ to $+z$ direction. Conjugate gradients method, $k_o = 1.0$.

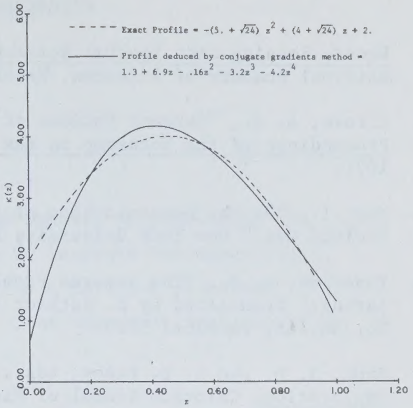


Figure 8. Profile of dielectric constant $\kappa(z) = -(5. + \sqrt{24})z^2 + (4 + \sqrt{24})\ z + 2$ for a slab with a perfect conductor a $z = 0$. Plane wave was incident from $+z$ to $-z$ direction. Conjugate gradients method, $k_o = 1$.

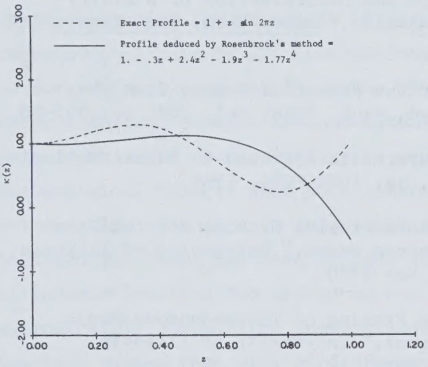


Figure 9 Profile of dielectric constant $\kappa(z) = 1 + z \sin 2\pi z$ for a dielectric slab. Rosenbrock's method, $k_o = 1$.

VI. CONCLUSIONS

Three methods of solving the remote probing problem were presented. These methods were developed especially to utilize data obtained in the angular spectrum domain. The first method presented a novel technique for extracting an exact linear integral equation from a nonlinear integral equation. The second method was based upon a new representation of the solution to the differential equation describing the problem. Numerically, the second method is more stable than the first.

The third method was seen to be numerical in nature and so lacks many of the analytical niceties which the first two methods possess. However, this method can use data from either the frequency domain or the angular spectrum domain and it can be used for geometries other than planar. Furthermore, this method is numerically stable.

REFERENCES

1. Remote Sensing with Special Reference to Agriculture and Forestry, National Academy of Sciences, Washington, D. C., 1970.
2. Kliore, A. J., "Current Methods of Radio Ossultation Data Inversion," Proceedings of the Workshop on the Mathematics of Profile Inversion, 1971.
3. Kay, I., "On the Determination of the Free Electron Distribution of an Ionized Gas," New York University Res. Rept. EM-141, September 1959.
4. Faddeyev, L. D., "The Inverse Problem in the Quantum Theory of Scattering," Translated by B. Seckler in New York University Res. Rept. No. EM-165, December 1960.
5. Haun, J. D. and L. U. LeRoy, eds., Subsurface Geology in Petroleum Exploration, Colorado School of Mines, Golden, Colorado, 1958.
6. Kelly, E., ed., Ultrasonic Energy, University of Illinois Press, Urbana, Illinois, 1965.
7. Gelfand, I. M. and B. M. Levitan, "On the Determination of a Differential Equation from its Spectral Function," Amer. Math. Soc. Trans., Series 2, Vol. 1, pp. 253-304, 1955.
8. Marchenko, V. A., "Reconstruction of the Potential Energy from the Phase of Scattered Waves," Dakl. Akad. Nauk. SSSR, Vol. 104, pp. 695-98.
9. Becher, W. D. and C. B. Sharpe, "A Synthesis Approach to Magnetotelluric Exploration," Radio Science, Vol. 4, pp. 1089-1094, 1969.
10. Schaubert, D. H. and R. Mittra, "Remote Probing Methods for the Determination of the Profile of Inhomogeneous Media," University of Illinois, Antenna Lab., Scientific Rept. #15, May 1970.
11. Mostafavi, M. and R. Mittra, "Remote Probing of Inhomogeneous Media Using Parameter Optimization Techniques," University of Illinois, Antenna Lab., Scientific Rept. #14, April 1970.
12. Sharpe, C. B., "The Synthesis of Infinite Lines," Quar. of Appl. Math., Vol. 21, p. 105, 1963.
13. Rosenbrock, H. H., "An Automatic Method for Finding the Greatest or Least Value of a Function," Computer Jour., October 1960.
14. Fletcher, R. and C. M. Reeves, "Function Minimization by Conjugate Gradients," Computer Jour., Vol. 7, No. 2, 1964.

BACKGROUND AND HISTORY OF PARALLEL PROCESSING

DANIEL L. SLOTNICK
University of Illinois
Urbana, Illinois 61801

ABSTRACT

The concept of parallel processing as opposed to pipelining and overlapping is presented as a stratagem to increase the operational speed of a digital computer. ILLIAC IV is a parallel processor patterned after the SOLOMON computers. Some important features of ILLIAC IV are discussed briefly.

The papers in this session all deal with the ILLIAC IV Computer. The ILLIAC IV Project was started in the Computer Science Department at the University of Illinois with the objective of developing a digital system employing the principle of parallel operation to achieve a computational rate of 10^9 instructions per second. In order to achieve this rate the system was to employ 256 processors operating simultaneously under a central control divided into four subassembly quadrants of 64 processors each. Due primarily to subcontractor problems several basic technological changes were necessitated during the course of the program, principally, reduction in individual logic circuit complexity and memory technology. These resulted in cost escalation and schedule delays ultimately limiting the system to one quadrant with an overall speed of approximately 200 million instructions per second. It is this one quadrant system that will be discussed for the remainder of this paper.

Over the past ten years designers have introduced a number of variations on the strictly sequential mode of operation. One stratagem has been to overlap the operation of the central processing unit and the operation of input-output devices. By means of a fine-grained separation of the computer's functional units a high degree of overlapping has been attained. Current efforts in "pipelining" the processing of "operands" will allow a further significant increase in speed.

Overlapping and pipelining, however, are both fundamentally limited in the advances in speed they can provide. The approach taken in ILLIAC IV surmounts fundamental limitations in ultimate computer speed by allowing--at least in principle--an unlimited number computational events to take place simultaneously. The logical design of ILLIAC IV is patterned after that of the SOLOMON computers, prototypes of which were built by the Westinghouse Electric Corporation in the early 1960's. In this design a single master control unit sends instructions to a sizable number of independent processing elements and transmits addresses to individual memory units associated with these processing elements ("processing-element memories"). Thus, while a single sequence of instructions (the program) still does the controlling, it controls a number of processing elements that execute the same instruction simultaneously on data that can be, and usually are, different in the memory of each processing element.

Each of the 64 processing elements of ILLIAC IV is a powerful computing unit in its own right. It can perform a wide range of arithmetical operations on numbers that are 64 binary digits long. These numbers can be in any one of six possible format; the number can be processed as a single number 64 bits long in either a fixed or a "floating" point representation, or the 64 bits can be broken up into smaller numbers of equal length. Each of the memory units has a capacity of 2,048 64-bit numbers. The time required to extract a number from memory (the access time) is 188 nanoseconds, but because additional logical circuitry is needed to resolve conflicts when two or more sections of ILLIAC IV call on memory simultaneously, the minimum time between successive operations of memory is increased to 350 nanoseconds.

Each processing element has more than 100,000 distinct electronic components assembled into some 12,000 switching circuits. A processing element together with its memory unit and associated logic is called a processing unit. In a system containing more than six million components one can expect a component or a connection to fail once every few hours. For this reason much attention has been devoted to testing and diagnostic procedures. Each of the 64 processing units will be subjected regularly to an extensive library of automatic tests. If a unit should fail one of these tests, it can be quickly unplugged and replaced by a spare, with only a brief loss of operating time. When the defective unit has been taken out of service, the precise cause of the failure will be determined by a separate diagnostic computer. Once the fault has been found and repaired the unit will be returned to the inventory of spares.

ILLIAC IV could not have been designed at all without much help from other computers. Two medium-sized Burroughs 5500 computers worked almost full time for two years preparing the artwork for the system's printed circuit boards and developing diagnostic and testing programs for the system's logic and hardware. These formidable design, programming and operating efforts were under the direction of Arthur B. Carroll, who during this period, was the project's deputy principal investigator.

The ILLIAC IV system is scheduled for completion by the end of this calendar year; the fabrication phase is essentially complete with some final assembly and considerable debugging yet to be completed.

In the following papers, Denenberg describes the logical organization of the Hardware, Randal describes the software required to make this hardware a computer system, Sameh deals with some of the unique questions arising in the efficient application of two parallel array algorithms, while McIntyre discusses some overall application questions and details the system's current status. The final paper by Mr. Bouknight is a description of the operating environment for this system, namely the ARPA Network of which the ILLIAC IV will be a principal node. All of this work was sponsored under a grant (Contract number USAF 30(602)4144) from the Advanced Research Projects Agency.

REFERENCES

1. The Solomon Computer, Daniel L. Slotnick, W. Carl Borck and Robert C. McReynolds, Proceedings of Fall Joint Computer Conference, 1962, pp. 97-107.

2. Unconventional Systems, Daniel L. Slotnick, Proceedings of Spring Joint Computer Conference, 1967, pp. 477-481.

ILLIAC IV HARDWARE STRUCTURE

STEWART A. DENENBERG
Center for Advanced Computation
University of Illinois, Urbana, Illinois

ABSTRACT

The architecture or hardware structure of the ILLIAC IV System is discussed. The ILLIAC IV System comprises the ILLIAC IV Array plus the ILLIAC IV I/O System. The ILLIAC IV Array is a Vector or Array Processor with a specialized Control Unit that can be viewed as a small stand-alone computer by itself.

A. ILLIAC IV in Brief

The original design of ILLIAC IV contained four Control Units: each of which controlled a 64 Arithmetic and Logic Unit (ALU) Array Processor. The version being built by the Burroughs Corporation will have only one Control Unit which drives 64 ALUs as shown in Figure 1. It is for this reason that ILLIAC IV is sometimes referred to as a Quadrant (one-fourth of the original machine) and it is this abbreviated version of ILLIAC IV that will be discussed for the remainder of this article.

One difference between ILLIAC IV and a general Array Processor is that the Control Unit (CU) has been decoupled from the rest of the Array Processor so that certain instructions can be executed completely within the resources of the CU at the same time that the ALU is performing its vector operations. In this way another degree of parallelism is exploited in addition to the inherent parallelism of 64 ALUs being driven simultaneously. What we have is 2 computers inside ILLIAC IV; one that operates on scalars and one that operates on vectors. All of the instructions however, emanate from the computer that operates on scalars--the CU.

Each element of the ALU Array is not called by its generic name (ALU) but is called a Processing Element or PE. There are 64 PEs and they are numbered from 0 to 63. Each PE responds to appropriate instructions if the PE is in an active mode. (There exist instructions in the repertoire which can activate or de-activate a PE.) Each PE performs the same operation under command from the CU in the lock-stepped manner of an Array Processor. That is, since there is only one Control Unit, there is only one instruction stream and all of the ALUs respond together or are lock-stepped to the current instruction. If the current instruction is ADD for example, then all of the ALUs will Add--there can be no instruction which will cause some of the ALUs to be adding while others are multiplying. Every ALU in the Array performs the instruction operation in this lock-stepped fashion, but the operands are vectors whose components can be and usually are different.

Each PE has a full complement of arithmetic and logical circuitry and under command from the CU will perform an instruction "at-a-crack" as an Array Processor. Each PE has its own 2048 word 64-bit memory called a Processing Element Memory (PEM) which can be accessed in about 350 ns. Special routing instructions can be used to move data from PEM to PEM. Additionally, operands can be sent to the PEs from the CU via a full-word (64 bit) one-way communication line and the CU has eight-word one-way communication with the PEM array (for instruction and data fetching).

An ILLIAC IV word is 64 bits and data numbers can be represented in either 64-bit floating point, 64-bit logical, 48-bit fixed point, 32-bit floating point, 24-bit fixed point, or 8-bit fixed point (character) mode. By utilizing the 64-bit, 32-bit and 8-bit data formats the 64 PEs can hold a vector of operands with either 64, 128, or 512 components. Since ILLIAC IV can add 512 operands in the 8 bit integer mode in about 66 nanoseconds, it is capable of performing almost 10^{10} of these "short" additions per second. ILLIAC IV can perform approximately 150 million 64-bit, rounded, normalized floating-point additions per second.

The I/O is handled by a B6500 Computer System. The Operating System, including the assemblers and compilers, also reside in the B6500.

B. The ILLIAC IV System

The ILLIAC IV System can be organized as in Figure 2. The ILLIAC IV System consists of the ILLIAC IV Array plus the ILLIAC IV I/O System. The ILLIAC IV Array consists of the Array Processor and the Control Unit. In turn, the Array Processor is made up of 64 Processing Elements (PEs) and their 64 associated memories--Processing Element Memories (PEMs). The ILLIAC IV I/O System is comprised of the I/O Subsystem, the Disk File System and the B6500 control computer. The I/O Subsystem is broken down further to the CDC, BIOM and IOS. The B6500 is actually a medium-scale computer system by itself.

The ILLIAC IV Array will be discussed first, in a general manner, followed by a brief description of the ILLIAC IV I/O System.

1. The ILLIAC IV Array

Figure 3 represents the ILLIAC IV Array--the Control Unit plus the Array Processor.

a. Control Unit (CU)

The Control Unit is not just the Control Unit that we're used to thinking of on a conventional computer but can be viewed as a small unsophisticated computer in its own right. Not only does it cause the 64 Processing Elements to respond to instructions, there is a repertoire of instructions that can be completely executed within the resources of the Control Unit, and the execution of these instructions is overlapped with the execution of the instructions which drive the Processing Element Array. Again, it is worthwhile to view ILLIAC IV as being two computers, one which operates on scalars and one which operates on vectors.

The Control Unit contains 64 integrated circuit registers called the ADVAST Data Buffer (ADB) which can be used as a high speed scratch-pad memory. ADVAST is an acronym for Advanced Station and is one of the five functional sections of the CU. Each register of the ADB (D0 through D63) is 64-bits long. The CU also has 4 Accumulator Registers called ACAR0, ACAR1, ACAR2, and ACAR3 each of which is also 64 bits long. The ACARs can be used as accumulators for integer addition, shifting, Boolean operations and holding loop control information--such as the lower limit, increment and upper limit. In addition the ACARs can be used as index registers to modify storage references within the memory section (PEM).

b. Processing Element (PE)

Each Processing Element (PE) is a sophisticated ALU capable of a wide range of arithmetic and logical operations. There are 64 PEs numbered 0 through 63. Each PE in the array has 6 programmable registers: the A register (RGA) or Accumulator, the B register (RGB) which holds the second operand in a binary operation (such as Add, Subtract, Multiply or Divide), the R or routing register (RGR) which transmits information from one PE to another, the S register (RGS) which can be used as temporary storage by the programmer, the X register (RGX) or index register to modify the address field of an instruction, and the D or mode register (RGD) which controls the active or nonactive status of each PE independently. The mode register determines whether a PE will be active or passive during instruction execution. Since this register is under the programmer's control, individual PEs within the array of 64 PEs may be set to enabled (active) or disabled (passive) status based on the contents of one of the other PE registers. For example, there are instructions which disable all PEs whose RGR contents are greater than their RGA contents. Only those PEs in an enabled state are able to execute the current instruction. All registers are 64 bits except RGX which is 16 bits and RGD which is 8 bits.

c. Processing Element Memory (PEM)

Each PE has its own 2048 word, 64-bits per word, random access memory. Each memory is called a Processing Element Memory or PEM and they are numbered 0 through 63 also. A PE and PEM taken together is called a Processing Unit or PU. PE_i may only access PEM_i so that one PU cannot modify the memory of another PU. Information can, however, be passed from one PU to another via the Routing Network which is one of the 4 paths by which data flows through the ILLIAC IV Array.

d. Data Paths

Besides the Instruction Control Path which drives the 64 PEs during the execution of an instruction there are four paths by which data flows through the ILLIAC IV Array. These paths are called the Control Unit Bus (CU Bus), the Common Data Bus (CDB), the Routing Network, and the Mode Bit Line.

i. Control Unit Bus (CU Bus)

Instructions or data from the PEMs in blocks of eight words can be sent to the CU via the Control Unit Bus (CU Bus). The instructions to be executed are distributed throughout the PEMs and are fetched in blocks of eight words to the CU via the CU Bus as necessary. Although the Operating System takes care of fetching and executing instructions, data can also be fetched in blocks of 8 words under program control using the CU Bus.

ii. Common Data Bus (CDB)

Information stored in the Control Unit can be "broadcast" to the entire 64 PE Array simultaneously via the Common Data Bus (CDB). A value such as a constant to be used as a multiplier need not be stored 64 times in each PEM; instead this value can be stored within a CU register and then broadcast to each enabled PE in the array. In addition the operand or address portion of an instruction is sent to the PE array via the CDB.

iii. Routing Network

Information in one PE register can be sent to another PE register by special routing instructions. (Information can be transferred from PE register to PEM by standard LOAD or STORE instructions.) High speed routing lines run between every RGR of every PE and its nearest left and right neighbor (distances of -1 and +1 respectively) and its neighbor 8 positions to the left and 8 positions to the right (-8 and +8 respectively). Other routing distances are effected by combinations of routing -1, +1, -8, or +8 PEMs; that is, if a route of 5 to the right is desired, the software will figure out that the fastest way to do this is by a right route of 8 followed by three left routes of 1. Figure 4 shows one way to view the connectivity which exists between PEs. As can be seen from the figure, PE_0 is connected to PE_{56}, PE_1, PE_8, and PE_{63}.

iv. Mode Bit Line

The Mode Bit Line consists of one line coming from the RGD of each PE in the Array. The Mode Bit Line can transmit one of the eight mode bits of each RGD in the array up to an ACAR in the Control Unit. If this bit is the bit which indicates whether or not a PE is on or off, we can transmit a "mode pattern" to an ACAR. This mode pattern reflects the status or on-offness of each PE in the array; then there are instructions which are executed completely within the Control Unit that can test this mode pattern and branch on a zero or non-zero condition. In this way branching in the instruction stream can occur based on the mode pattern of the entire 64 PE array.

2. ILLIAC IV Input/Output (I/O) System

The ILLIAC IV Array is an extremely powerful information processor, but it has of itself no I/O capability. The I/O capability along with the supervisory system (including compilers and utilities) reside within the ILLIAC IV I/O System. The ILLIAC IV I/O System (see Figure 5) consists of the I/O Subsystem, a Disk File System (DFS) and a B6500 Control Computer (which in turn supervises a large Laser Memory, a Data Communications Processor, and the ARPA Network Link). The total ILLIAC IV System consisting of the ILLIAC IV I/O System and the ILLIAC IV Array is shown in Figure 6. All system configurations shown are transitory, and more than likely will have changed several times in the next year or so.

a. I/O Subsystem

The I/O Subsystem consists of the Control Descriptor Controller (CDC), the Buffer Input/Output Memory (BIOM) and the Input/Output Switch (IOS).

i. Control Descriptor Controller (CDC)

The CDC monitors a section of the CU waiting for an I/O request to appear. The CDC can then interrupt the B6500 Control Computer which can, in turn, try to honor the request and place a response code back in that section of the CU via the CDC. This response code indicates the status of the I/O request to the program in the ILLIAC IV Array.

The CDC causes the B6500 to initiate the loading of the PE Memory Array with programs and data from the ILLIAC IV Disk (also called the Disk File System or DFS). After PE Memory has been loaded, the CDC can then pass control to the CU to begin execution of the ILLIAC IV Program.

ii. Buffer Input/Output Memory (BIOM)

The B6500 Control Computer can transfer information from its memory through its CPU at the rate of 80×10^6 bits/second. The ILLIAC IV Disk (DFS) accepts information at the rate of 500×10^6 bits/second. This factor of over six in information transfer rates between the two systems necessitates the placing of a rate-smoothing buffer between them. The BIOM is that buffer. A buffer is also necessary for the conversion of 48-bit B6500 words to 64-bit ILLIAC IV words which can come out of the BIOM two at a time via the 128 bit wide path to the Disk File System. The BIOM is actually four PE memories providing 8192 words of 64-bit storage.

iii. Input/Output Switch (IOS)

The IOS performs two functions. As its name implies, it is a switch and is responsible for switching information from either the Disk File System or from a port which can accept input from a real time device. All bulk data transfers to and from the PE Memory Array are via IOS. As a switch it must insure that only one input is sending to the Array at a given time. In addition, the IOS acts as a buffer between the Disk File System and the Array, since each channel from the ILLIAC IV Disk to the IOS is 256 bits wide and the bus from the IOS to the PE Memory Array is 1024 bits wide.

b. Disk File System (DFS)

The Disk File System (DFS) consists of two Storage Units, two Electronics Units and two Disk File Controllers. The DFS is also called the ILLIAC IV Disk or simply, the Disk. The Disk is of 10^9-bit capacity, having 128 heads, with one head per track. The DFS has two channels, each of which can transmit or receive data at a rate of $.5 \times 10^9$ bits/second over a path 256 bits wide; however, if both channels are sending or receiving simultaneously the transfer rate is 10^9 bits/second.

c. B6500 Control Computer

The B6500 Control Computer consists of a Central Processing Unit (CPU), Memory, a Multiplexor and a set of Peripheral Devices (Card Reader, Card Punch, Line Printer, 4 Magnetic Tape Units, 2 Disk Files and Console Printer and Keyboard). It is the function of the B6500 to manage all programmers' requests for system resources. This means that the Operating System will reside on the B6500. All compiling and assembling of programs is also performed on the B6500. Utilities, such as Card-to-Disk, Card-to-Tape, etc. are also executed on the B6500. From a total System standpoint, the ILLIAC IV Array can be considered as a special-purpose peripheral device of the B6500 capable of solving certain classes of problems with extremely high speed.

i. Data Communications Processor

The Data Communications Processor will supervise a set of remote terminals. The terminals are devices such as Teletypes or CRT displays that allow the user access to the ILLIAC IV System. Users will be able to enter their jobs into the system in either a batch or interactive mode via the terminals. The terminals can also be used to monitor jobs while in execution and to scan the ILLIAC IV Disk or Laser Memory which will contain the output from a job. If the user decides he needs a hard copy of his output, he can then signal the system from his terminal to activate a printer.

ii. Laser Memory

The B6500 supervises a 10^{12}-bit read-only Laser Memory developed by the Precision Instrument Company. The beam from an argon laser records binary data by burning microscopic holes in a thin film of metal coated on a strip of polyester sheet, which is carried by a rotating drum. Each data strip can store some 2.9 billion bits. A "strip file" provides storage for 400 data strips containing more than a trillion bits. The time to locate data stored on any one of the 400 strips is five seconds. Within the same strip data can be located in 200 milliseconds. The read and record rate is four million bits a second on each of two channels. A projected use of this memory will allow the user to "dump" large quantities of programs and data into this storage medium for leisurely review at a later time; hard copy output can optionally be made from files within the Laser Memory.

iii. ARPA Network Link

The ARPA Network is a group of computer installations separated geographically but connected by high speed (50,000 bits/second) data communication lines. On these lines, the members of the "Net" can transmit information--usually in the form of programs, data, or messages. The link performs an information switching function and is handled by an IMP (Interface Message Processor) and a Network Control Program stored within each member installation's "host" computer. Each IMP operates in a "store and forward mode", that is, information in one IMP is not lost until the receiving IMP has signalled complete reception and retention of the message. The IMP interfaces with each member's computer system and converts information into standard format for transmission to the rest of the Net. Conversely, the IMP accepts information in a standard format and converts it to the particular data format of the member installation. In this way, the ARPA Network is a form of a computer utility with each contributing member offering its unique resources to all of the other members.

REFERENCES

1. D. L. Slotnick, "The Fastest Computer", Scientific American, Vol. 224, No. 2, February 1971, pp. 76-87.

2. Stewart A. Denenberg, "An Introductory Description of the ILLIAC IV System", ILLIAC IV Document No. 225, Department of Computer Science File No. 850, Urbana, Ill.: Center for Advanced Computation, University of Illinois, July 15, 1971.

3. G. H. Barnes, et al., "The ILLIAC IV Computer", IEEE Transactions on Computers, Vol. C-17, No. 8, August 1968, pp. 746-757.

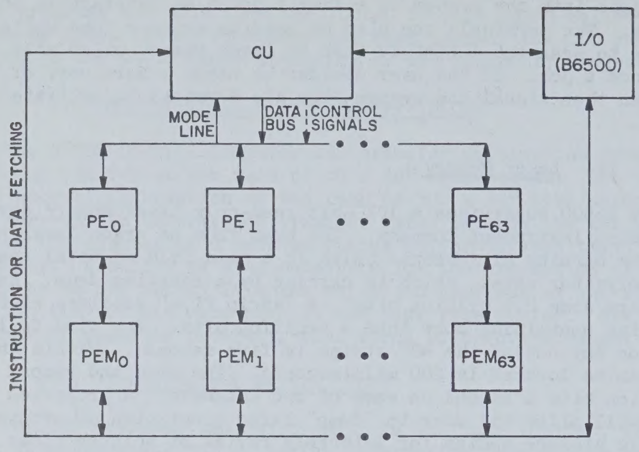

CU
I/O
(B6500)
MODE
LINE
DATA
BUS
CONTROL
SIGNALS
INSTRUCTION OR DATA FETCHING
PE0
PE1
PE63
PEM0
PEM1
PEM63

Figure 1. Functional Block Diagram of ILLIAC IV

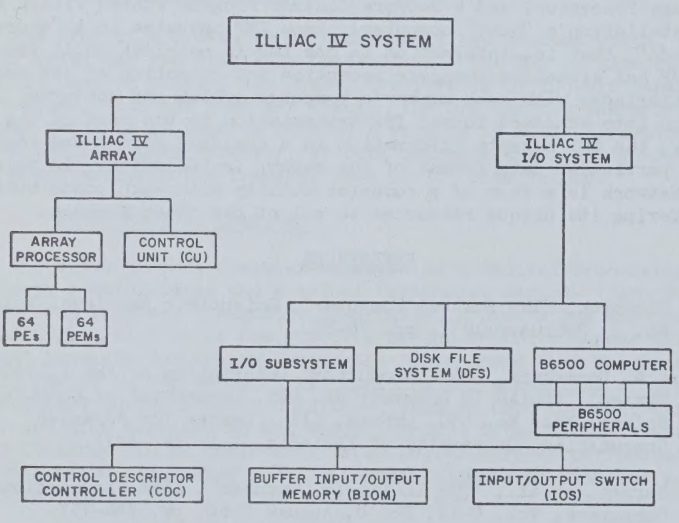

ILLIAC IV SYSTEM
ILLIAC IV
ARRAY
ILLIAC IV
I/O SYSTEM
ARRAY
PROCESSOR
CONTROL
UNIT (CU)
64
PEs
64
PEMs
I/O SUBSYSTEM
DISK FILE
SYSTEM (DFS)
B6500 COMPUTER
B6500
PERIPHERALS
CONTROL DESCRIPTOR
CONTROLLER (CDC)
BUFFER INPUT/OUTPUT
MEMORY (BIOM)
INPUT/OUTPUT SWITCH
(IOS)

Figure 2. ILLIAC IV System Organization

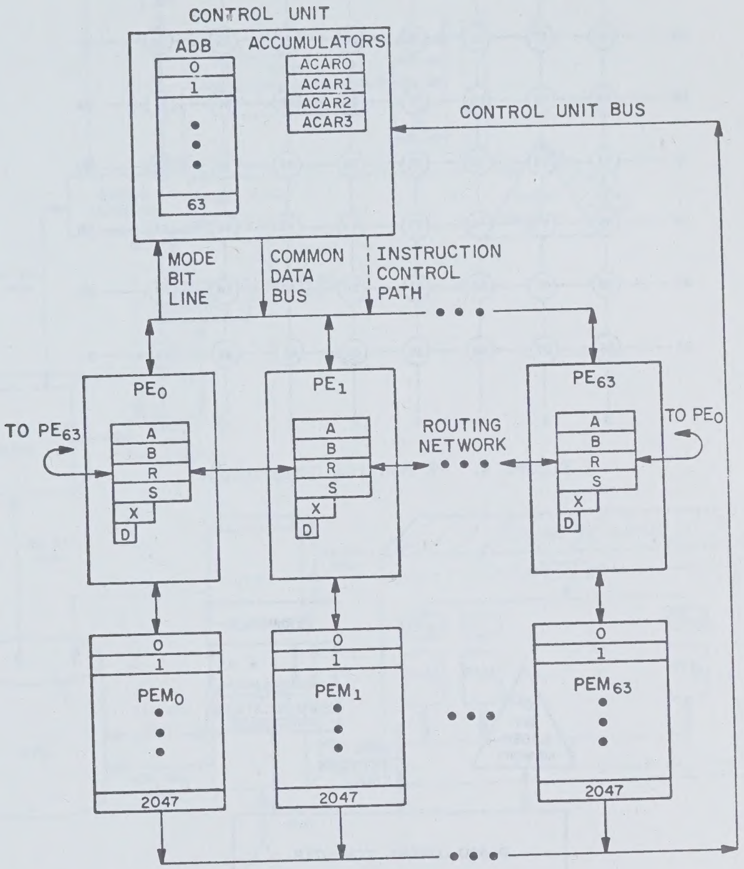

CONTROL UNIT
ADB
0
1
63
ACCUMULATORS
ACAR0
ACAR1
ACAR2
ACAR3
CONTROL UNIT BUS
MODE BIT LINE
COMMON DATA BUS
INSTRUCTION CONTROL PATH
PE$_0$
PE$_1$
PE$_{63}$
A
B
R
S
X
D
TO PE$_{63}$
TO PE$_0$
ROUTING NETWORK
PEM$_0$
PEM$_1$
PEM$_{63}$
0
1
2047

Figure 3. ILLIAC IV Array

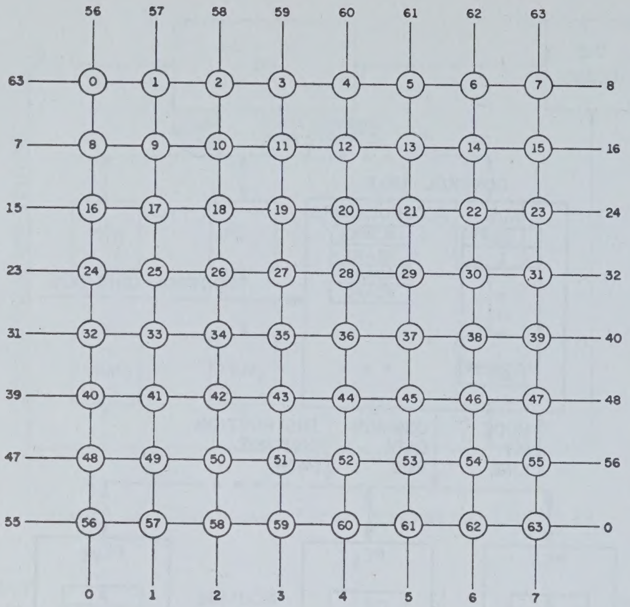

56
57
58
59
60
61
62
63
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63

Figure 4. PE Routing Connections

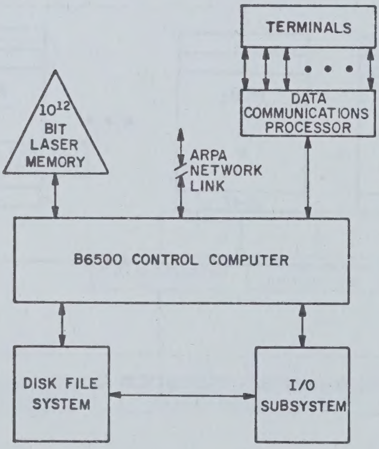

TERMINALS
DATA COMMUNICATIONS PROCESSOR
10^{12} BIT LASER MEMORY
ARPA NETWORK LINK
B6500 CONTROL COMPUTER
DISK FILE SYSTEM
I/O SUBSYSTEM

Figure 5. ILLIAC IV I/O System

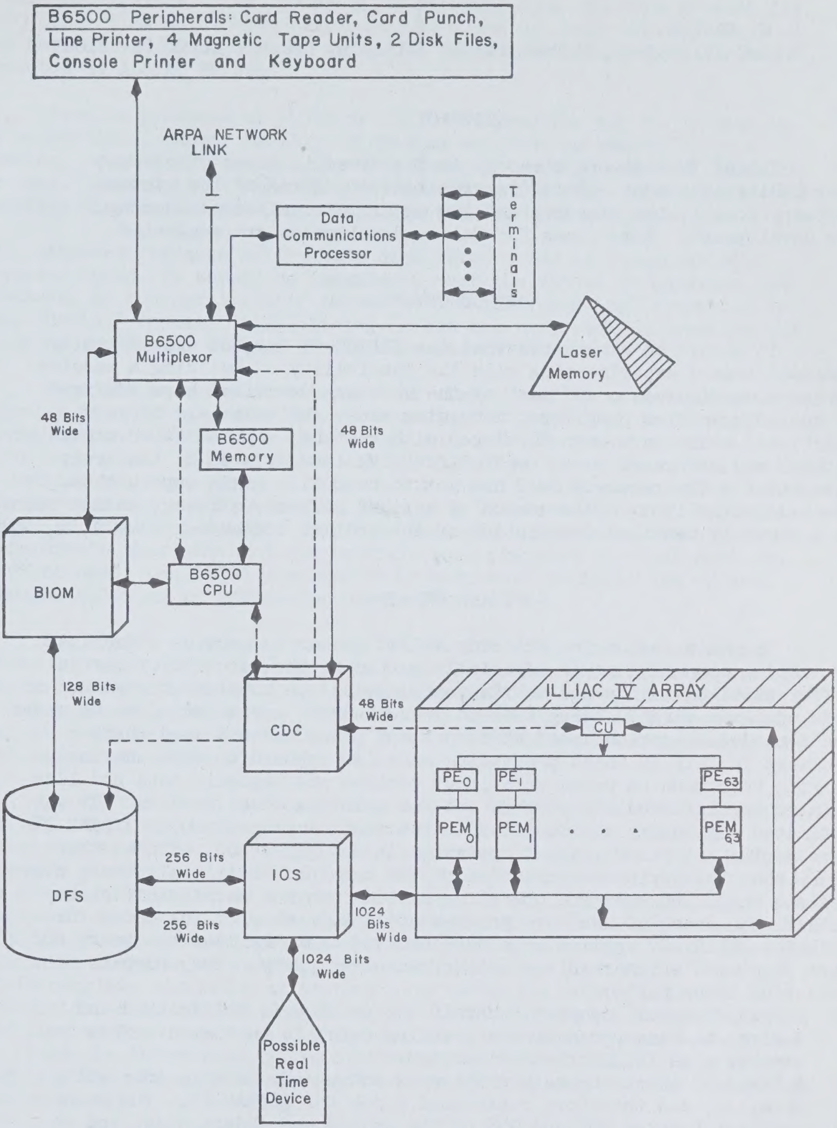

B6500 Peripherals: Card Reader, Card Punch, Line Printer, 4 Magnetic Tape Units, 2 Disk Files, Console Printer and Keyboard
ARPA NETWORK LINK
Data Communications Processor
Terminals
B6500 Multiplexor
Laser Memory
48 Bits Wide
B6500 Memory
48 Bits Wide
BIOM
B6500 CPU
128 Bits Wide
CDC
48 Bits Wide
ILLIAC IV ARRAY
CU
PE0
PE1
PE63
PEM0
PEM1
PEM63
DFS
256 Bits Wide
IOS
256 Bits Wide
1024 Bits Wide
1024 Bits Wide
Possible Real Time Device

Figure 6. ILLIAC IV System

ILLIAC IV SOFTWARE: AN OVERVIEW

J. M. RANDAL
ILLIAC IV Project, University of Illinois, Urbana, Illinois 61801

ABSTRACT

ILLIAC IV Software strategy is described in terms of current user habits and needs. Brief descriptions are given of the systems software itself, its history, and the major lessons learned during its development. Some ideas for future development are suggested.

INTRODUCTION

It should be remembered that the ILLIAC IV Project was initially directed toward experimenting with the feasibility of building a massive hardware configuration and most of the software described here (defined as operating system compilers, debugging aids, and necessary library functions) could have been developed with suitably sophisticated simulators, without any reference to a "real" ILLIAC IV; indeed, most of the truly innovative software envisioned has yet to be built. This paper, then, is devoted primarily to a discussion of a sound software strategy rather than to a minutely detailed description of the initial software. [See 1, 2, 3]

SOFTWARE STRATEGY

The main reason for building ILLIAC IV was to provide a facility of massive computing power especially suited to the solution of partial differential equations and matrix manipulation. A considerable amount of this work was already being done on less powerful serial machines by users who demanded execution speed at almost any price and who used their machines largely in batch processing mode (as opposed to time sharing), relying very much on punch card input devices and magnetic tape and line printer for intermediate storage and for printing final results. It was estimated that users would utilize a substantial proportion of ILLIAC IV time during its first year of operation in debugging and refining their code, thus justifying the creation of the machine. Nearly all these users desired that languages and operating systems provide machine efficiency rather than "ease of use" or "programmability." Many of the codes for existing ILLIAC IV applications have evolved over the last ten years and have displayed a remarkable architectural similarity. They comprise the following three parts:

1) A "Preprocessor" section, wherein the problem is initialized and decimal-to-binary conversion (reading data) is performed. This is usually a serially oriented section.
2) A "Kernel" where the main problem is addressed, usually inherently parallel, and therefore considered a job for ILLIAC IV. The Kernel occupies between 80% and 95% of the source and object code, and on serial machines uses a similar proportion of CPU time.
3) A "Postprocessor" section, wherein results are stored on archive files, necessary binary-to-decimal (writing data) conversions are made, graphs are plotted, and line printer output formatting is set. This is usually a serially oriented section. From here control may loop back to the Kernel in order to complete a further set of iterations on the data.

Generally speaking, ILLIAC IV jobs will be presented to the B6700 as card decks, tape files, or as files received over the ARPA Network [4]. B6700 disk files which have originated from one of these sources, but have been edited through local or remote on-line consoles, may also be presented as ILLIAC IV jobs.

Results produced by B6500 or ILLIAC IV programs may be printed in the conventional manner locally, displayed on local or remote on-line consoles, or transmitted over the ARPA Network to output devices local to the user. Later, it is expected that provision for microfilm graphics and selective viewing and editing of results will be made available.

However, because of the high disk latency (40 μs) compared with processor speed, it should be remembered that the ILLIAC IV hardware, as it stands, is not particularly amenable to a "time-sharing" operation if "time-sharing" implies "time-slicing." But all the usual interactive and debugging facilities will be provided on the B6500.

From the intended users' point of view, then, it seemed adequate initially to provide simple batch processing software that would enable jobs to be run efficiently, even though only one job kernel would be active at a time. This attitude was reenforced by questions of reliability. The initial mean time between failure for the ILLIAC IV array was envisioned to be between two and four hours. Debugging "state-of-the-art" software on untried "state-of-the-art" hardware demanded either more bravery or more foolhardiness than most ordinary mortals were prepared to volunteer. On the other hand, the B6700 promised to be much more reliable and offered Burroughs ALGOL as an implementation language.

The wisest course was clearly to aim for simple batch processing systems software executed as much as possible on the B6700. As experience leads to greater familiarity and confidence with the array hardware and initial software performance, those modules whose operation is inherently parallel could be moved onto ILLIAC IV if the B6700 became overloaded, or if it seemed otherwise advisable to do so.

One of the first software tasks was to write an ILLIAC IV simulator that would handle all operations in the CU and PEs and all ILLIAC IV machine instructions. Although this simulator ran approximately a million times more slowly than ILLIAC IV on a B5500, and later, about 200,000 times more slowly on a B6700, it became the basis for all language and algorithm development for ILLIAC IV. Recently, the simulator was extended to simulate ILLIAC IV I/O subsystem logic and its interaction with the ILLIAC Operating System. At the time of writing, potential users may "run ILLIAC IV jobs" on this complete simulator as if they were using the complete ILLIAC IV installation, although the length of the ILLIAC IV program must, of course, be of limited execution time. This simulated Operating System is built so that it may be incrementally transferred onto the real ILLIAC IV hardware. Hopefully, the present software will be operational very shortly after the hardware is built and working.

It is also worth noting that all ILLIAC IV languages except the Assembler, but including the ILLIAC Control Language, have been implemented using a compiler compiler system called TWST [5]. This has allowed a certain amount of language experimentation, yet has resulted in the early availability of usable languages despite the fact that the team providing the basic ILLIAC IV software included never more than fourteen professionals. This relatively small team has provided a coherent set of software, although the relatively mechanized approach has caused some degradation in compile time speed, for instance. Consequently, there is a considerable amount of refinement to be done. However, this will not affect the users, whose programs will not have to be changed while the overall compiler efficiency increases.

OPERATING SYSTEM

The ILLIAC IV Operating System [6] operates in a "diagnostic" or "normal" mode. The main task of the diagnostic mode is the testing and diagnosis of possible faults in the ILLIAC IV I/O subsystem and the ILLIAC IV array itself. These diagnostic programs are designed to identify faults and to automatically identify the pluggable unit in which they occur. The unit is then replaced and the Operating System automatically re-runs the test program that identified the fault to ensure that it has in fact been remedied. While in this mode, an interactive routine is available to the engineers to enable them to either call specific diagnostic programs or generate new ones. This system may also be used to interrogate ILLIAC IV registers and to change their values. While in the diagnostic mode, the B6700 is available to carry on its usual work of preparing jobs to be run on ILLIAC IV and to process the output from those which have already run. However, it cannot use any major unit that is being diagnosed at that time by the diagnostic routines. In the "normal" mode, the Operating System administers the running on ILLIAC IV and the use of the I/O subsystem in this mode.

The ILLIAC IV Operating System consists of a set of asyncronous processes which run under the control of the B6700 Master Control Program (MCP). When a user submits an ILLIAC IV job to the B6500, it usually consists of the following parts:

1) B6500 programs usually written in B6500 ALGOL or FORTRAN which transform and prepare binary input files (input under format control, i.e., character-to-binary conversion) to be used by the ILLIAC IV program called the "Preprocessor."
2) ILLIAC IV programs usually written in ASK, Glypnir, or ILLIAC IV FORTRAN which use ILLIAC IV to operate on the files prepared by the B6500 programs and to prepare binary output files ("Kernel").
3) B6500 programs usually written in B6500 ALGOL or FORTRAN which transform binary files from ILLIAC IV to the required external form for use or storage ("Postprocessor").
4) An ILLIAC Control Language (ICL) program which defines the job. The ILLIAC Control Language controls the Operating System for the job which it defines, and thus may be seen as "driving" the Operating System for that particular job, although the Operating System may be operating on several ICL programs concurrently.

As each job enters the system through the B6500, it is assigned a priority (if one has not been specified by the user). Higher priority jobs, as well as being processed more favorably by the B6700, have the privilege of having their ILLIAC IV program parts preempting running ILLIAC IV programs of lower priority. The preempting program will run to completion unless it is itself preempted by a program of higher priority. The number of stacked up preempted jobs may be arbitrarily set when the system is initialized. However, the main function of the preemption scheme is to allow debugging runs and short production runs to move along the job stream quite rapidly, while longer production runs and jobs of lower priority ensure that the ILLIAC IV itself is rarely idle for any length of time. If the work load permits, non-ILLIAC IV jobs may be run on the B6700 while it is not compiling, preprocessing, postprocessing and administering ILLIAC IV jobs.

B6700 programs and ILLIAC IV programs communicate via ILLIAC IV disk files (for data) and the 48-bit path through the CU and TMU (interrupt signals). The protocol for these signals over the 48-bit path is administered by two modules. The first is a small executive program residing in ILLIAC IV itself (called OS4) which processes all interrupts for the Array, handles all communications between the user program and the rest of the Operating System, and provides a few standard functions for use in the Array. OS4 communicates with a module (known as the Job Partner) in the B6500 which acts as a clearing house for all communication between OS4 and thus the user program running on ILLIAC IV. The Job Partner thus initiates all data transfers between the B6500 and ILLIAC IV, B6700 and ILLIAC IV Disk, and between ILLIAC IV and ILLIAC IV disk. This arrangement emphasizes the rate of the B6500 as an I/O computer for the ILLIAC IV or, conversely, the ILLIAC IV as a peripheral for the B6500.

The ILLIAC IV Control Language, ICL, is used to coordinate the execution and communication of the set of B6700 and ILLIAC IV programs that constitute an ILLIAC IV job. ICL is an ALGOL-like block structured language in which FILEs, PROGRAMs, and INTEGERs may be declared.

FILEs may be declared in formats and number representations appropriate to either machine. Additionally, ILLIAC IV files may be laid out on ILLIAC IV disk in a very flexible manner to allow the user to minimize disk latency. When files are moved from one machine to another, i.e., when ILLIAC IV files are equated to B6700 files, they are automatically transformed to the appropriate format and binary representation.

PROGRAMs may be declared as ILLIAC IV programs, B6500 compilers, Job Partners, ICL programs, or other B6700 programs. A program may be viewed as acting much like an INTEGER PROCEDURE with monstrous side effects. The files which the program uses for data and results and the file on which the program is to be found are also included in the declaration.

INTEGERs are used mainly for manipulating a single integer result always delivered by a program when it is called by the ICL program. This integer may be set by the user, and normally indicates the degree of success of the program execution.

The availability of IF and CASE statements allows considerable flexibility in job control, and facilities are provided to allow limited simultaneous execution of B6700 and ILLIAC IV programs within the same job.

The ILLIAC Control Language allows the user to construct a simple Operating System appropriate to each of his jobs. The ILLIAC IV Operating System then allows each of these "job operating systems" to overlap each other in their utilization of time and resources. It also encourages the division of the job into managable modules, thus making it more resiliant to hardware and software malfunction.

The present Operating System can undergo considerable development with the present hardware configuration. Apart from simply enhancing the existing code and making it part of the Burroughs MCP, many of the file conversion modules and the disk allocator, whose operations are inherently parallel, may be moved to the ILLIAC IV array. Preparation for this work is well underway. Major changes in operating philosophy are best left until hardware performance has been evaluated in the real world.

LANGUAGES

The strategy adopted in language development for ILLIAC IV was more liberal than that adopted for the Operating System. Apart from the Assembler, three "higher level languages" were attempted. Two survived to be usable. The problem of defining requirements for the languages was more difficult than that of the Operating System, where user habits and needs were fairly well understood.

The first major difficulty is that of formal representation. FORTRAN, ALGOL, and other numerically oriented higher level languages are based primarily upon traditional mathematical formalism. However, there is no comparably suitable formalism that can be easily used as a basis for describing the kind of operations that ILLIAC IV does easily and, at the same time, can be extended to a truly non-trivial set of array or matrix operations. For while operations like vector and matrix operations such as addition and multiplications by a scalar quantity are trivial to implement, matrix multiplication and inversion require reasonably sized subroutines. In addition, in the field of partial differential equations the large number of treatments of boundary conditions, discontinuities, and other special cases have presented a variety of approaches that has only recently been attacked from the higher level language point of view [7].

The second problem is that ILLIAC IV can be fiendishly difficult to program properly if one does not banish nearly all serial machine preconceptions and habits. Realizing this, the language designer is faced with choosing between completely disguising the architecture of ILLIAC IV, thus helping the programmer by doing quite a lot of dirty housekeeping for him, or fearlessly exposing the architecture of ILLIAC IV, thus, in all probability, forcing the programmer to re-think his algorithm in ILLIAC IV terms. The first approach was applied in the design of the Tranquil programming language, the second in Glypnir. The nature and success of each project will be discussed later in this paper.

The third problem was the unexpected amount of code that had to be produced for any given problem, and the difficulty of optimizing this code automatically. A study of ILLIAC IV Assembler codes indicates that for every arithmetical operation (including fetch and store), at least two others were required to direct and define the scope of the arithmetic instructions. About half of these extra instructions were CU instructions setting up loops, constants, administering ADB storage, and storing stacks of mode patterns. The other half were concerned with mode pattern calculation. Thus, it was difficult to see what efficient code generation meant for ILLIAC IV, and to make allowance for it, especially in view of the fact that Tranquil and Glypnir were being implemented before the ILLIAC IV instruction code had been finalized.

In this latter respect, it may be interesting to note that all compilers were designed to translate source code into ILLIAC IV Assembler code. This will make the compilers themselves more resilient in the face of machine modification, and also, since neither of the higher level languages do any automatic optimization, give the enthusiastic user a chance to optimize his own code should he wish to do so.

The ILLIAC IV Assembler [8] is fairly conventional as Assemblers go nowadays. It has a very sophisticated macro definition facility which may be used to include standard I/O facilities and other communications with OS4. It also has "pseudo operations" to help with storage layout and PE address allocation, of which there are four types: syllable (half-word) addresses, word addresses, row addresses, and I/O word addresses. All of these may be used by some instructions.

Before embarking upon a more detailed discussion of the ILLIAC IV higher level languages, it is worth pointing out that their differences with existing languages center around the following subjects:

- Storage allocations in a two dimensional store. The "natural" method of addressing PE memory is by row (or 64 words). Single words of PE memory may be addressed individually. However, a "column", that is, a group of words in the memory of a single PE, may not be addressed as a group.
- The expression of parallelism and mode control. To a certain extent, a vector may be the natural expression of parallelism, and ILLIAC IV languages should in some sense be able to allow operations on vectors or rows of matrices. The length of the vector and the elements within it to be operated upon are defined by the mode pattern. ILLIAC IV languages should follow efficient and comprehensive calculation and manipulation of mode patterns.
- The expression of routing and indexing. The language should allow reasonable expression of routing, and of indexing by a different amount in each PE.

The first ILLIAC IV language attempted was "Tranquil" [9], an ALGOL-like language entirely independent of machine organization. It was designed to allow programmers to operate on arrays of data in a simultaneous way. An algorithm may be viewed as being applied to any set of elements of an array at the same logical time, thus often allowing a programmer to avoid thinking about such details as how to index through an array and which elements to save in temporary arrays.

The Tranquil compiler was brought to a fairly advanced state of completion when experience with the object code indicated that the overhead involved in completely masking the machine's architecture was too high in relation to the users' demand for fast execution of their programs. This was not just a matter of the payoff between the cost of programming time and execution time, but of making existing programs, whose execution times for proposed problems were unfeasibly long on existing equipment, run in a realistic time on the ILLIAC IV. Users were thus prepared to do a certain amount of reprogramming in, say, FORTRAN and to restructure their algorithms to suit ILLIAC IV architecture, but they were unwilling to reprogram the whole problem in ILLIAC IV Assembler.

As it seemed unlikely that Tranquil could be made sufficiently efficient before the machine was then projected to be delivered, work on Tranquil was halted, and the job of providing an extended FORTRAN was begun. However, Tranquil should not be shrugged off. It represents the level of language that implementors would like to provide. Continued experience with Extended FORTRAN has shown, moreover, that the inefficies of Tranquil could have been remedied with time, and that with slight modification to the language, Tranquil could have become a viable and extremely interesting and useful ILLIAC IV language.

Glypnir [10] is also an ALGOL-like block structured language and in many ways an ILLIAC IV equivalent of Burroughs B6700 ALGOL in the sense that it was written to allow the knowledgeable user to exploit ILLIAC IV architecture to considerable advantage.

Program data types or procedures may be declared as being either CU or PE variables. In the first case, they refer to one 64-bit value corresponding to an ILLIAC IV word, and in the second case, they refer to a row of ILLIAC IV words and thus may have up to 64 different values simultaneously. Boolean values, however, are neither CU or PE variables but refer to 64 true-false values. Thus, a boolean variable is said to be TRUE when all its constituent bits are TRUE and FALSE when all its constituent bits are FALSE. Two extra operators, SOME and EVERY, are used to provide a bridge between word and bit level logic.

All arithmetic operations are carried out under the control of a MODE pattern. The MODE allows the 64 true-false values of a boolean to be associated with each of the PEs. When a bit of the MODE pattern is TRUE, the corresponding PE is enabled and may thus deliver the results of an operation.

The Glypnir expression A:=B*C; where A, B, and C are PE variables means that each value of A is multiplied by C and delivered to A for each enabled PE. If C were a CU variable, it would be repeated 64 times in an invisible PE variable before the statement was evaluated.

Special facilities exist to allow the rotation and shifting of rows to the right and left, in a way similar to the more familiar operations conventionally carried out on words, thus allowing the "route" instruction to be utilized.

FOR and IF statements are also provided, but often give unconventional results. For instance, given PE variables A, B, and C, the statement

```
IF A > B THEN C:=A ELSE C:=B;
```

will deliver the maximum elements of A and B to C, and may result in both the THEN and ELSE statements being executed.

Glypnir will allow the explicit inclusion of blocks of Assembler language for the optimization of any section of code, and also has facilities within the language to refer explicitly to selected hardware registers for those who wish to optimize without going into Assembler.

The language has now been in use with the ILLIAC IV simulator for over a year and a considerable amount of useful experience has been gained from it. Projected future developments include the addition of a 32-bit mode and the facility to allow binding to separately compiled Assembler, Glypnir, or Cockroach subroutines.

ILLIAC IV FORTRAN [11] was provided more for the general ILLIAC IV user than for the specialist. Glypnir demands that the programmer undertake the detailed supervision of storage allocation and be constrained to think in terms of ILLIAC IV rows or vectors of ILLIAC IV rows. ILLIAC IV FORTRAN allows the user to think in terms of rows of any length, and to think of arrays in terms of "straight" format where each column is staggered across PEs. This latter devise allows rows and columns to be accessed with equal facility.

In order to provide mode control, the data type BINARY has been added to the language which allows arbitrary arrays of bits (or true-false values) to be declared in a way similar to arrays of LOGICAL values.

The DO statement has been extended to allow parallel execution of arithmetic expressions, and extra constructs have been added to the language to allow the shifting and rotation of vectors and array rows.

The only significant change in definition applies to EQUIVALENCE and COMMON statements where the two dimensional store of ILLIAC IV imposes restrictions on the usual serial definition. However, as these restrictions are those usually forced upon the programmer by "real" problems, they may often be a help rather than a hindrance.

One very important option allowed in ILLIAC IV FORTRAN is that of taking an existing serial FORTRAN code with suitably adjusted I/O statements and running it on ILLIAC IV serially (in one PE) to test its validity before incrementally parallelizing it.

An ILLIAC IV FORTRAN to Glypnir translator has now been in use for several months and is considered to be very successful. ILLIAC IV FORTRAN exhibits many of the operational features of Tranquil, and has benefitted from both the Tranquil and the Glypnir experiences.

Because of the large disk latency compared with processing speed, the I/O software for ILLIAC IV languages tends to be elaborate in order to enable the programmer to synchronize his I/O functions with his computing. A considerable amount of disk latency may be removed by using the elaborate disk layout facilities provided by the operating system. This latency may be reduced further by the I/O intrinsics in both Glypnir and ILLIAC IV FORTRAN.

File declarations are modeled on those of the Burroughs B6700, with the option of implicit buffering to a level given by the user. The user is also given the option of driving his program by I/O interrupt. At a lower level, standard assembler macros provide a similar facility for the assembler programmer.

In terms of what could be done, and, by analogy, has been done in producing software for some serial machines, ILLIAC IV software development is still very much in its infancy. While the present software is more than adequate to get a considerable amount of useful work out of ILLIAC IV for a few years after its completion, there are several areas where new work would produce valuable payoffs.

ILLIAC IV offers hardware facilities for "32-bit mode," that is, the ability to store and operate two 32-bit numbers of differing value stored in one ILLIAC IV word. None of the present compilers offer 32-bit mode as a complete data type that can be used in harmony with 64-bit mode. Work in this area is continuing at the University, and consists not only of inserting the necessary language constructions into the compilers and developing the necessary standard algorithms, but also bringing 32-bit user experience to some sort of parity with 64-bit user experience. The effective use of 32-bit mode may essentially double the power of ILLIAC IV, but much remains to be done before patterns of effective use are established.

The automatic overlaying of array and program segments is also particularly important in view of the relatively small core storage capacity of the ILLIAC IV. This, of course, impinges upon the Operating System as well as the compilers, and although automatic overlay has always been desirable, its detailed investigation has always, quite rightly, been postponed until sufficient experience with existing hardware sheds more light on its feasibility. At present, the same effect may be gained by coupling program modules together with a suitable ILLIAC Control program, or by explicitly overlaying file segments into array space.

Lastly, neither of the compilers optimize code in the sense that they scramble instructions to make the best use of CU-PE overlap. Considerable savings are envisioned with this kind of optimization but, once again, this is difficult to do with a simulator which takes no account of this overlay.

CONCLUSION

All the software described in this paper, and much more that is not, has been tried by users through the agency of simulation. The major task of transferring this software to the real machine still lies ahead, but all the software, especially the languages, has gone through several stages of appraisal and revision in the light of informed user comment. It is hoped that this relatively conservative software policy will provide a usable ILLIAC IV very soon after the hardware is completed, and a firm base for a more adventurous second generation of software.

REFERENCES

1. McIntyre, D. E.,"An Introduction to the ILLIAC IV Computer," DATAMATION, April 1970.
2. Barnes, G. H., et al., "The ILLIAC IV Computer," IEEE Trans. Comp., Vol. C-17, No. 8, August 1968, pp. 746-757.
3. Kuck, D. J., "ILLIAC IV Software and Applications Programming," IEEE Trans. Comp., Vol. C-17, No. 8, August 1968, pp. 758-770.
4. Roberts, L.G., and Wessler, B.D., Computer Network Development to Achieve Resource Sharing, AFIPS Conference Proceedings, May 1970, pp. 543-549.
5. Trout, H. R. G., A BNF-Like Language for the Description of Syntax Directed Compilers (M.S. Thesis), ILLIAC IV Document No. 161, January 13, 1969.
6. ILLIAC IV Software Reference Manual, Vol. 1, Chapter 2, ILLIAC IV Project, University of Illinois at Urbana-Champaign, Urbana, Ill.
7. Gary, J., PDELAN--A Programming Language for the Solution of Partial Differential Equations, ILLIAC IV Document No. 229, July 16, 1970.
8. ILLIAC IV Software Reference Manual, Vol. 2, Chapter 1, ILLIAC IV Project, University of Illinois at Urbana-Champaign, Urbana, Ill.
9. Abel, N. E. et al., Tranquil, A Language for an Array Processing Computer, AFIPS Conference Proceedings, May 1969.
10. ILLIAC IV Software Reference Manual, Vol. 2, Chapter 2, ILLIAC IV Project, University of Illinois at Urbana-Champaign, Urbana, Ill.

11. ILLIAC IV Software Reference Manual, Vol. 2, Chapter 3, ILLIAC IV Project, University of Illinois at Urbana-Champaign, Urbana, Ill.

ILLIAC IV APPLICATIONS

AHMED H. SAMEH
Center for Advanced Computation, University of Illinois
Urbana, Illinois

ABSTRACT

Two difficulties face the users of parallel machines. First, the numerical algorithms must be modified to suit the array nature of the arithmetic units. Second, the data may have to be stored in highly interleaved memories so that proper combinations of elements are available simultaneously at various steps of the algorithm. In this paper we discuss the implementation on the ILLIAC IV of two algorithms for finding the eigenvalues and eigenvectors of real symmetric matrices, and an algorithm for the solution of linear systems of equations using Householder's transformations.

1. INTRODUCTION

Most of the numerical algorithms available on conventional machines cannot be readily modified for efficient parallel computations. Two difficulties face the users of parallel machines. First, algorithms must be devised that are suited to the array nature of their arithmetic units. Second, data may have to be stored in highly interleaved memories so that proper combinations of elements are available simultaneously at various steps of the algorithm. In this paper we discuss the following examples, Jacobi and Householder algorithms for finding the eigenvalues and eigenvectors of real symmetric matrices, and reducing a real nonsymmetric matrix to the upper-triangular form using Householder's transformations.

2. JACOBI'S METHOD

In the classical method of Jacobi [1], a real symmetric matrix can be reduced to a diagonal matrix by a sequence of plane rotations,

$$A_{k+1} = R_k A_k R_k^t \qquad (k=1,2,\ldots.) \tag{1}$$

where $A_1 = A$ is the $n\times n$ original matrix, and each rotation R_k differs from the identity matrix only in the elements,

$$R_{pp}^{(k)} = R_{qq}^{(k)} = \cos\theta_{pq}^{(k)}; \quad R_{pq}^{(k)} = -R_{qp} = \sin\theta_{pq}^{(k)}, \quad (p<q) \tag{2}$$

By properly choosing the angle $\theta_{pq}^{(k)}$, the off-diagonal elements $a_{pq}^{(k+1)} = a_{qp}^{(k+1)}$ of A_{k+1} in (1) are eliminated. It can be easily shown that, [1], $\theta_{pq}^{(k)}$ is given by,

$$\cos^2\theta_{pq}^{(k)} = \tfrac{1}{2}(1 + \frac{x_k}{y_k}); \quad \sin^2\theta_{pq}^{(k)} = \tfrac{1}{2}(1 - \frac{x_k}{y_k}) \tag{3}$$

where, $x_k = |a_{pp}^{(k)} - a_{qq}^{(k)}|$, $y_k = (4a_{pq}^2 + x_k^2)^{1/2}$. Restricting $\theta_{pq}^{(k)}$ by $|\theta_{pq}^{(k)}| \le \pi/4$, then $\cos\theta_{pq}^{(k)}$ will always be taken positive and $\sin\theta_{pq}^{(k)}$ will be of the same sign as $[2a_{pq}^{(k)}/(a_{pp}^{(k)} - a_{qq}^{(k)})]$. From the definition (2) of the rotations R_k we can see that only the rows and columns p and q of A_k are affected by the transformation (1). Clearly the implementation of such an algorithm on a parallel machine would be highly inefficient. For, even if updating of the $2(n-1)$ elements (A_{k+1} is symmetric and $a_{pq}^{(k+1)} \equiv 0$ hence need not be computed) is performed with all the PE's working; it requires only one or two PE's at the most to compute $R_{pp}^{(k)}$, $R_{pq}^{(k)}$.

It is possible, however, to modify Jacobi's algorithm so as to eliminate more than one off-diagonal element (above the main diagonal) per transformation, and hence make the method more suited for parallel computations [2]. For example, for a matrix A of an even order $n = 4$, if the orthogonal transformation R_1 is chosen as,

$$\begin{bmatrix} R_{11}^{(1)} & R_{12}^{(1)} & 0 & 0 \\ -R_{12}^{(1)} & R_{22}^{(1)} & 0 & 0 \\ 0 & 0 & R_{33}^{(1)} & R_{34}^{(1)} \\ 0 & 0 & -R_{34}^{(1)} & R_{44}^{(1)} \end{bmatrix}$$

where $R_{pq}^{(1)}$ for all the pairs (p,q) are given by (2), then $\theta_{12}^{(1)}$ and $\theta_{34}^{(1)}$ can be independently chosen such that the elements $a_{12}^{(2)}$ and $a_{34}^{(2)}$ of $A_2 = R_1 A_1 R_1^t$ are annihilated. Similarly, by appropriately choosing the elements of,

$$R_2 = \begin{bmatrix} R_{11}^{(2)} & 0 & R_{13}^{(2)} & 0 \\ 0 & R_{22}^{(2)} & 0 & R_{24}^{(2)} \\ -R_{13}^{(2)} & 0 & R_{33}^{(2)} & 0 \\ 0 & -R_{24}^{(2)} & 0 & R_{44}^{(2)} \end{bmatrix}, \quad R_3 = \begin{bmatrix} R_{11}^{(3)} & 0 & 0 & R_{14}^{(3)} \\ 0 & R_{22}^{(3)} & R_{23}^{(3)} & 0 \\ 0 & -R_{23}^{(3)} & R_{33}^{(3)} & 0 \\ -R_{14}^{(3)} & 0 & 0 & R_{44}^{(3)} \end{bmatrix}$$

the elements $a_{13}^{(3)}$, $a_{24}^{(3)}$ of A_3, and $a_{14}^{(4)}$, $a_{23}^{(4)}$ of A_4 are eliminated respectively. Consequently, after $(n-1)$ transformations each of $1/2\ n(n-1)$ off-diagonal element (above the main diagonal) has been eliminated once, each transformation eliminating $n/2$ such elements. Let each $(n-1)$ orthogonal transformations be denoted by a sweep, then for our specific example $(n=4)$ the second sweep will consist of the orthogonal matrices R_4, R_5, and R_6 which have the same construction as R_1, R_2, and R_3 respectively, and so on. Thus, for matrices of even order (see [2] for the general case), the elements of each of the $(n-1)$ orthogonal matrices R_k within each sweep are given by,

$$\begin{aligned} R_{pp}^{(k)} &= R_{qq}^{(k)} = \cos\theta_{pq}^{(k)} \\ R_{pq}^{(k)} &= -R_{qp}^{(k)} = \begin{cases} \sin\theta_{pq}^{(k)} & p<q \\ -\sin\theta_{pq}^{(k)} & p>q \end{cases} \end{aligned} \tag{4}$$

where p and q are sequences defined by,

(a) for $k = 1,2,3,\ldots\ldots, \frac{n}{2} - 1$

$$q = \frac{n}{2} - k + 1,\ \frac{n}{2} - k + 2,\ \ldots\ldots,\ n-k$$

$$p = \begin{cases} (n - 2k + 1) - q & \frac{n}{2} - k + 1 \le q \le n - 2k \\ 2(n - k) - q & n - 2k < q \le n - k - 1 \\ n & n - k - 1 < q \end{cases} \tag{5}$$

(b) for $k = \frac{n}{2}, \frac{n}{2} + 1, \ldots\ldots, n - 1$

$$q = n - k,\ n - k + 1,\ \ldots\ldots,\ \frac{3}{2}n - k - 1$$

$$p = \begin{cases} n & q < n - k + 1 \\ 2(n - k) - q & n - k + 1 \le q \le 2(n - k) - 1 \\ (3n - 2k - 1) - q & 2(n - 1) < q \end{cases} \tag{6}$$

the remaining elements of R_k are, of course, zero. For each k the angles $\theta_{pq}^{(k)}$ are determined such that the elements $a_{pq}^{(k)}$ are annihilated for all the pairs (p,q). For example, for $n = 8$ and $k = 4$ the pairs (p,q) are determined by (6) $\{(8,4); (3,5); (2,6); (1,7)\}$, and R_4 is of the form

$$\begin{bmatrix} R_{11}^{(4)} & & & & & & R_{17}^{(4)} & \\ & R_{22}^{(4)} & & & & R_{26}^{(4)} & & \\ & & R_{33}^{(4)} & & R_{34}^{(4)} & & & \\ & & & R_{44}^{(4)} & & & & R_{48}^{(4)} \\ & & -R_{35}^{(4)} & & R_{55}^{(4)} & & & \\ & -R_{26}^{(4)} & & & & R_{66}^{(4)} & & \\ -R_{17}^{(4)} & & & & & & R_{77}^{(4)} & \\ & & & -R_{48}^{(4)} & & & & R_{88}^{(4)} \end{bmatrix} \tag{7}$$

Therefore, denoting each element eliminated, in a given sweep, in the k-th transformation by the integer k, the pattern of the annihilated elements for a matrix of order 16 is shown below,

$$\begin{array}{cccccccccccccccc}
\ast & 7 & \textcircled{14} & 6 & \textcircled{13} & 5 & \textcircled{12} & 4 & \textcircled{11} & 3 & \textcircled{10} & 2 & \textcircled{9} & 1 & \textcircled{8} & \textcircled{15} \\
 & \ast & 6 & \textcircled{13} & 5 & \textcircled{12} & 4 & \textcircled{11} & 3 & \textcircled{10} & 2 & \textcircled{9} & 1 & \textcircled{8} & \textcircled{15} & \textcircled{14} \\
 & & \ast & 5 & \textcircled{12} & 4 & \textcircled{11} & 3 & \textcircled{10} & 2 & \textcircled{9} & 1 & \textcircled{8} & \textcircled{15} & 7 & \textcircled{13} \\
 & & & \ast & 4 & \textcircled{11} & 3 & \textcircled{10} & 2 & \textcircled{9} & 1 & \textcircled{8} & \textcircled{15} & 7 & \textcircled{14} & \textcircled{12} \\
 & & & & \ast & 3 & \textcircled{10} & 2 & \textcircled{9} & 1 & \textcircled{8} & \textcircled{15} & 7 & \textcircled{14} & 6 & \textcircled{11} \\
 & & & & & \ast & 2 & \textcircled{9} & 1 & \textcircled{8} & \textcircled{15} & 7 & \textcircled{14} & 6 & \textcircled{13} & \textcircled{10} \\
 & & & & & & \ast & 1 & \textcircled{8} & \textcircled{15} & 7 & \textcircled{14} & 6 & \textcircled{13} & 5 & \textcircled{9} \\
 & & & & & & & \ast & \textcircled{15} & 7 & \textcircled{14} & 6 & \textcircled{13} & 5 & \textcircled{12} & \textcircled{8} \\
 & & & & & & & & \ast & \textcircled{14} & 6 & \textcircled{13} & 5 & \textcircled{12} & 4 & 7 \\
 & & & & & & & & & \ast & \textcircled{13} & 5 & \textcircled{12} & 4 & \textcircled{11} & 6 \\
 & & & & & & & & & & \ast & \textcircled{12} & 4 & \textcircled{11} & 3 & 5 \\
 & & & & & & & & & & & \ast & \textcircled{11} & 3 & \textcircled{10} & 4 \\
 & & & & & & & & & & & & \ast & \textcircled{10} & 2 & 3 \\
 & & & & & & & & & & & & & \ast & \textcircled{9} & 2 \\
 & & & & & & & & & & & & & & \ast & 1 \\
 & & & & & & & & & & & & & & & \ast
\end{array}$$

Let us consider now the implementation of this modified scheme on a parallel machine. For the sake of simplicity of the illustrations we will assume that the size of the matrix is equal to the number of PE's, and furthermore we will ignore the fact that the matrix is symmetric and hence we need only operate on those elements on and above the main diagonal. Since each row of the matrix $B = R_k A_k$ is obtained by summing multiples of two rows of A_k, and each column of $A_{k+1} = B R_k^t$ is obtained by summing multiples of two columns of B. Therefore, in order to minimize the number of idle PE's in the matrix multiplication $R_k A_k R_k^t$, it is essential to store the elements of A_k such that each row or column can be accessed with one memory fetch. This is achieved by storing A_k in the "skewed" form. In such a storage scheme we map the array elements into PE memory as follows,

$$a_{ij} \rightarrow \text{loc. } i, \qquad \text{PE } (i + j - 1)(\text{mod } N) \tag{8}$$

where N is the number of PE's being used. For example, for N = 8 the elements of an 8×8 matrix are stored as follows,

LOC.	PE1	PE2	PE3	PE4	PE5	PE6	PE7	PE8
1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8
2	2,8	2,1	2,2	2,3	2,4	2,5	2,6	2,7
3	3,7	3,8	3,1	3,2	3,3	3,4	3,5	3,6
4	4,6	4,7	4,8	4,1	4,2	4,3	4,4	4,5
5	5,5	5,6	5,7	5,8	5,1	5,2	5,3	5,4
6	6,4	6,5	6,6	6,7	6,8	6,1	6,2	6,3
7	7,3	7,4	7,5	7,6	7,7	7,8	7,1	7,2
8	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,1

PE1	PE2	PE3	PE4	PE5	PE6	PE7	PE8
$a_{11}^{(4)}$	$a_{17}^{(4)}$	$a_{22}^{(4)}$	$a_{26}^{(4)}$	$a_{33}^{(4)}$	$a_{35}^{(4)}$	$a_{44}^{(4)}$	$a_{48}^{(4)}$
$a_{77}^{(4)}$		$a_{66}^{(4)}$		$a_{55}^{(4)}$		$a_{88}^{(4)}$	

Let us consider the transformation $A_5 = R_4 A_4 R_4^t$ where R_4 is given by (7). By routing, the following configuration can be obtained, however, some of the PE's have to remain idle since the routing distances for all the elements are not the same. Now, from (3) $\sin\theta_{pq}^{(5)}$ and $\cos\theta_{pq}^{(5)}$ for the pairs (p,q): (1,7); (2,6); (3,5); (4,8) are computed, simultaneously with roughly half the PE's being idle. Once all the elements of R_4 are computed, the matrix multiplication $R_4 A_4 R_4^t$ is performed with all the PE's working. For example, to compute the first row of $B = R_4 A_4$,

(i) broadcast $R_{11}^{(4)}$ and $R_{17}^{(4)}$ to all PE's.

(ii) fetch the first row of A_4 and multiply all its elements by $R_{11}^{(4)}$,

(iii) fetch the seventh row of A_4, multiply all its elements by $R_{17}^{(4)}$, route all the elements of the resultant vector a distance of +2, and add to (ii)

This process is repeated until all the rows are computed. Storing the matrix B in the skewed form, each column of A_5 is obtained as follows,

(i) broadcast $-R_{48}^{(4)}$ and $R_{88}^{(4)}$ to all PE's

(ii) fetch the fourth column of B, multiply all its elements by $-R_{48}^{(4)}$, and then route them a distance of +4 (iii)fetch the eighth column of B, multiply all its elements by $R_{88}^{(4)}$, add to (ii) and store in the same locations of the eighth column of A_4.

Assuming that the matrix A has converged, (using some criterion [1]), to the diagonal form after u sweeps or after r orthogonal transformations, where $r = u(n - 1)$, then the diagonal elements of $A_{r+1} = WAW^t$ are taken to be the eigenvalues. The columns of $W^t = V_1^t V_2^t \ldots V_u^t$ are the corresponding eigenvectors, where for the j-th sweep,

$$V_j^t = (R_1^t)_j (R_2^t)_j \ldots (R_{n-1}^t)_j .$$

3. HOUSEHOLDER'S METHOD

This method [1,3] reduces a real symmetric matrix $A=[a_{ij}],(i,j=1,..,n)$, to the tridiagonal matrix $T=[t_{ij}]$, ($t_{ij}=0$ for $|i-j|>1$), using elementary Hermitian matrices. There are (n-2) steps in this reduction. We define the matrices A_r by the relation $A_r = P_r A_{r-1} P_r$ $(r=1,2,\ldots,n-2)$ (9)

where $A_0 = A$ is the original matrix and the matrices P_r are defined by,

$$P_r = I - (u_r u_r^t)/2K_r^2 \tag{10}$$

in which u_r^t is a row vector whose first r elements are zero,

$$u_r^t = (0,\ldots,0,a_{r,r+1} \pm S_r, a_{r,r+2},\ldots,a_{r,n})$$

$$S_r = [\sum_{i=r+1}^{n} a_{ri}^2]^{1/2} \tag{11}$$

$$2K_r^2 = S_r^2 \pm a_{r,r+1} S_r$$

and the elements a_{ij} in (11) are those of A_{r-1}. The sign in both expressions $(a_{r,r+1} \pm S_r)$ and $(S_r^2 \pm a_{r,r+1} S_r)$ is chosen such that

$$|a_{r,r+1} \pm S_r| = |a_{r,r+1}| + S_r \tag{12}$$

So, in the r-th step, $(A_r = P_r A_{r-1} P_r)$, $(n - r - 1)$ zeros are introduced in the r-th row and column of A_{r-1} without destroying the zeros introduced in the previous step. Taking advantage of symmetry, the matrices A_r may be defined by the relations,

$$A_r = A_{r-1} - q_r u_r^t - u_r q_r^t \tag{13}$$

where u_r is given by (11), and

$$q_r = p_r - (u_r u_r^t p_r)/4K_r^2 \tag{14}$$

in which,

$$p_r = (A_{r-1} u_r)/2K_r^2 \tag{15}$$

The total number of multiplication required to reduce the matrix to the tridiagonal form is essentially $2/3\ n^3$.

It is clear that the major computations involved in each step are in computing p_r, hence we would like to store the matrix A_{r-1} such that we can access any column by one memory fetch. Therefore we can either store the matrix A in the skewed form described in the previous example or in the straight form where a_{ij} is in location i and PE j. Adopting this straight storage and assuming that n(the size of the matrix) = N(the number of PE's of the parallel machine) = 64, then the first step in the reduction can be described as follows. (1) The elements $(a_{1j}^{(o)})^2$, $(j \geq 2)$, are obtained by one multiplication and S_1^2 in (11) is obtained in $\log_2 64 = 6$ additions. Evaluating the square root S_1 and computing the scalar $\alpha_1 = (1/2\ K_1^2)$ is an inefficient process since only few PE's can be used. (2) The elements $a_{1j}^{(1)}$ $(j > 2)$ are set to zero, and $a_{12}^{(1)}$ is set to $\pm S_1$, while $a_{11}^{(1)} = a_{11}^{(o)}$. (3) The vector $\tilde{u}_1 = \alpha_1 u_1$ is trivially obtained by one multiplication and one addition. (4) Disable PE1, broadcast the first element $\tilde{u}_{1,1}$ of $\tilde{u}_1$ to PE's 2-64, fetch the second row of A_o $(a_{2j}^{(o)}$, j=2,3...,64), multiply its elements by $\tilde{u}_{1,1}$. (5) Broadcast the second element $\tilde{u}_{1,2}$ to PE's 2-64, fetch the third row of A_o (a_{3j}, j=2,3,...,64), multiply its elements by $\tilde{u}_{1,2}$, and add the resulting vector to that of (4). (6) Repeating the process for the rest of the rows of A_o, we obtain the non-zero elements of p_1. (7) Having the non-zero elements $\tilde{u}_{1,j}$ and $p_{1,j}$ in PE j, (j=2,3,...,64), then the scalar $\beta_1 = \tilde{u}_1^t\ p_1$ is obtained by one multiplication and $\log_2 64 = 6$ additions. (8) Broadcasting β_1 to PE's 2-64, multiplying all the elements of u_1 by $1/2\ \beta_1$, and subtracting the resulting vector from p_1, we obtain q_1 whose elements overwrite those of p_1. (9) Broadcasting the first element of q_1, $(q_{1,1})$, to PE's 2-64 and multiply all the elements of u_i by $q_{1,1}$. (10) Broadcast the first element of u_1, $(u_{1,1})$, to PE's 2-64, multiply all the elements of q_1 by $u_{1,1}$, and add the resulting vector to that of (9). (11) The second row of A_1 is obtained by subtracting (10) from the second row of A_o. (12) The process is repeated to obtain the rest of the rows of A_1.

It is clear that the number of idle PE's increase by one after each step and on the average 50% of the PE's are inactive during the tridiagonalization process. However such an inefficiency is not bad afterall, since the total number of operations in tridiagonalization ($2/3\ n^3$ multiplications on a serial machine) is negligible compared to that of finding all the eigenvalues and eigenvectors of A. The eigenvalues of the tridiagonal matrix T (and hence of A) lie in the interval [-b,b], where $b = ||T||_\infty = \max_i \sum_j |t_{ij}|$. Dividing this interval in 63 subintervals and for all μ_k (k=1, 2,...,64), where $\mu_1 = -b$, $\mu_{64} = +b$, the following quantities; $f_o(\mu_k)=1$, $f_1(\mu_k) = t_{11} - \mu_k$, and $f_i(\mu_k) = (t_{i,i} - \mu_k)\ f_{i-1}\ (\mu_k) - t_{i-1,i}^2\ f_{i-2}\ (\mu_k)$ i=2,3,...,64 are evaluated in parallel, one PE to each μ_k. For each μ_k the number of agreements in sign $s(\mu_k)$ of consecutive members of the sequence $f_i(\mu_k)$, $(i=0,1,\ldots,64)$, is the number of eigenvalues larger than μ_k. The process is repeated for those interval $[\mu_k, \mu_{k+1}]$ that contain more than one eigenvalue until all the eigenvalues are separated. Assuming that all the eigenvalues are distinct, then assigning each interval to a PE all the eigenvalues are evaluated in parallel by successive bisection.

Once all the eigenvalues are evaluated, the eigenvectors of the tridiagonal matrix T are obtained by the method of inverse iteration, each PE evaluating an eigenvector. Finally if z_i $(i=1,2,\ldots,64)$ is an eigenvector of T, then the corresponding eigenvector y_i of the original matrix A is computed by

$$y_i = P_1 P_2 \ldots P_{n-2} z_i$$

where $P_i = I - (u_i u_i^t/2K_i^2)$ and u_i $(i=1,2,\ldots,64)$ are assumed to be stored in the primary memory.

4. SOLUTION OF SYSTEMS OF LINEAR EQUATIONS

Let the linear system of equation be denoted by $Ax = b$, where A is an $n\times n$ real nonsymmetric matrix and b is a real n-dimensional vector. The matrix A can be reduced to an upper triangular matrix by a series of $(n-1)$ elementary Hermitian matrices. If we define the matrices A_r by the relations, $A_r = P_r A_{r-1}$ $(r=1,2,\ldots,n-1)$, where $A_o = A$ is the original matrix then A_{n-1} is an upper triangular matrix. The matrices P_r are given by (10) in which u_r^t is a row vector whose first $(r-1)$ elements are zero and is given by,

$$\begin{aligned} u_r^t &= (0,\ldots,0,a_{r,r} \pm S_r, a_{r+1,r},\ldots, a_{n,r}) \\ S_r &= \Big[\sum_{i=r}^{n} a_{i,r}^2 \Big]^{1/2} \\ 2K_r^2 &= S_r^2 \pm S_r \cdot a_{r,r} \end{aligned} \tag{16}$$

where the elements a_{ij} are those of A_{r-1}. Again we face the problem of operating on matrices that get smaller in size after each step. To overcome this problem while still having access to any row of A by one memory fetch it is highly desirable to store A such that we can have simultaneous access to each of the N elements of any $\sqrt{N} \times \sqrt{N}$ submatrix of A, where $N = 2^m$ (m is an even integer) is the number of PE's of the parallel machine. Thus during each step we can operate on any $\sqrt{N} \times \sqrt{N}$ submatrix, that does not contain eliminated elements, with the least number of PE's being idle. For n=N=16, (for example), the storage scheme is shown [4],

PE LOC.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1, 1	1, 2	1, 3	1, 4	2, 1	2, 2	2, 3	2, 4	3, 1	3, 2	3, 3	3, 4	4, 1	4, 2	4, 3	4, 4
2	4, 5	4, 6	4, 7	4, 8	1, 5	1, 6	1, 7	1, 8	2, 5	2, 6	2, 7	2, 8	3, 5	3, 6	3, 7	3, 8
3	3, 9	3,10	3,11	3,12	4, 9	4,10	4,11	4,12	1, 9	1,10	1,11	1,12	2, 9	2,10	2,11	2,12
4	2,13	2,14	2,15	2,16	3,13	3,14	3,15	3,16	4,13	4,14	4,15	4,16	1,13	1,14	1,15	1,16
5	8,14	5, 1	5, 2	5, 3	5, 4	6, 1	6, 2	6, 3	6, 4	7, 1	7, 2	7, 3	7, 4	8, 1	8, 2	8, 3
6	7, 8	8, 5	8, 6	8, 7	8, 8	5, 5	5, 6	5, 7	5, 8	6, 5	6, 6	6, 7	6, 8	7, 5	7, 6	7, 7
7	6,12	7, 9	7,10	7,11	7,12	8, 9	8,10	8,11	8,12	5, 9	5,10	5,11	5,12	6, 9	6,10	6,11
8	5,16	6,13	6,14	6,15	6,16	7,13	7,14	7,15	7,16	8,13	8,14	8,15	8,16	5,13	5,14	5,15
9	12, 3	12, 4	9, 1	9, 2	9, 3	9, 4	10, 1	10, 2	10, 3	10, 4	11, 1	11, 2	11, 3	11, 4	12, 1	12, 2
10	11, 7	11, 8	12, 5	12, 6	12, 7	12, 8	9, 5	9, 6	9, 7	9, 8	10, 5	10, 6	10, 7	10, 8	11, 5	11, 6
11	10,11	10,12	11, 9	11,10	11,11	11,12	12, 9	12,10	12,11	12,12	9, 9	9,10	9,11	9,12	10, 9	10,10
12	9,15	9,16	10,13	10,14	10,15	10,16	11,13	11,14	11,15	11,16	12,13	12,14	12,15	12,16	9,13	9,14
13	16, 2	16, 3	16, 4	13, 1	13, 2	13, 3	13, 4	14, 1	14, 2	14, 3	14, 4	15, 1	15, 2	15, 3	15, 4	16, 1
14	15, 6	15, 7	15, 8	16, 5	16, 6	16, 7	16, 8	13, 5	13, 6	13, 7	13, 8	14, 5	14, 6	14, 7	14 8	15, 5
15	14,10	14,11	14,12	15, 9	15,10	15,11	15,12	16, 9	16,10	16,11	16,12	13, 9	13,10	13,11	13,12	14, 9
16	13,14	13,15	13,16	14,13	14,14	14,15	14,16	15,13	15,14	15,15	15,16	16,13	16,14	16,15	16,16	13,13

This storage scheme maps an element a_{ij} into the memory as follows,

$$a_{ij} \to \text{loc. } [N^{1/2}\lfloor(i-1)/N^{1/2}\rfloor + \lfloor(j-1)/N^{1/2}\rfloor + 1];$$

$$\text{PE. } \{[((i-1) \bmod N^{1/2})N^{1/2} + \lfloor(i-1)/N^{1/2}\rfloor + j - 1] \pmod{N^{1/2}} + 1\} \quad (17)$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

Let us consider the r-th step in reducing A to an upper triangle. Denoting $(1/2\ K_r^2)$ by α_r then we have,

$$A_r = P_r A_{r-1} = A_{r-1} - \alpha_r u_r u_r^t A_{r-1} \quad (18)$$

In terms of the $\sqrt{N} \times \sqrt{N}$ submatrices $A_{k\ell}^{(r)}$, $(k,\ell=1,2,3,4)$, of A_r and the (4×1) subvectors $u_k^{(r)}$ of u_r we have,

$$A_{k\ell}^{(r)} = A_{k\ell}^{(r-1)} - \alpha_r u_k^{(r)} \sum_{j=1}^{4} u_j^{(r)^t} A_{j\ell}^{(r-1)} \quad (19)$$

Now for $r=4$ (say), the computations involved in obtaining A_{23} (for example) can be described as follows. (1) The scalar α_4 and the vector u_4 are trivially obtained as explained in the previous section. Note that the first three elements of u_4 are zero. (2) Denote $u_{4,i}$ to be the i-th element of u_4. Broadcast $u_{4,13}$ to PE's 12-15, $u_{4,14}$ to PE's 16, 1-4, $u_{4,15}$ to PE's 4-7, $u_{4,16}$ to PE's 8-11, fetch the elements in loc. 15, multiply the elements in each PE, and route the resulting elements a distance -1. (3) Broadcast $u_{4,9}$ to PE's 11-14, $u_{4,10}$ to PE's 15-16, 1-2, $u_{4,11}$ to PE's 3-6, $u_{4,12}$ to PE's 7-10, fetch the elements in loc. 11, multiply, add to the elements of (2), and route a distance -1. (4) Broadcast $u_{4,5}$ to PE's 10-13, $u_{4,6}$ to PE's 14-16, 1, $u_{4,7}$ to PE's 2-5, fetch the elements in loc. 7, multiply, add to the elements of (3).

(5) Enable only PE's 5-8, broadcast $u_{4,4}$ to those PE's, fetch the elements in loc. 3, multiply, route a distance +1, add to the elements in (4), and store the resulting elements in loc. 17. (6) Form the 16-dimensional vector with the elements $\alpha_4 u_{4,5}$ in PE's 10-13, $\alpha_4 u_{4,6}$ in PE's 14-16, 1, $\alpha_4 u_{4,7}$ in PE's 2-5, and $\alpha_4 u_{4,8}$ to PE's 6-9. (7) Fetch row 17 and multiply each of its elements with the corresponding elements of the vector in (6). (8) Fetch row 17, route its elements a distance of -4, multiply them by the corresponding elements of the vector in (6), and add to (7). (9) Fetch row 17, route a distance of -8, multiply by (6), add to (8). (10) Fetch row (17) again, route a distance of +4, multiply by (6), and add to (9). The final row of elements obtained are the elements of the submatrix $A_{23}^{(4)}$.

If we update the original vector $b = b_o$ after each step by, $b_r = b_{r-1} - \alpha_r u_r u_r^t b_{r-1}$ which is straightforward operation, then the vector x is computed by backsubstitution where the number of idle PE's increases by one after obtaining each x_i $(i=n,n-1,....,1)$.

REFERENCES

1. Wilkinson, J. H., The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.
2. Sameh, A. H., "On Jacobi and Jacobi-Like Algorithms for a Parallel Computer", Mathematics of Computation 25 (1971) 579-590.
3. Martin, R. S., et al. "Householder's Tridiagonalization of a Symmetric Matrix", Numer. Math. 11 (1968) 181-195.
4. Kuck, D. and Sameh, A. "Parallel Computation of Eigenvalues of Real Matrices", IFIP Congress 1971.

ILLIAC IV APPLICATIONS AND CURRENT STATUS

DAVID E. McINTYRE
ILLIAC IV Project
University of Illinois
Urbana, Illinois

Abstract

ILLIAC IV is uniquely suited to evaluate the function $f(\vec{x})$ simultaneously on up to 64 distinct argument sets $\vec{x}_i$. Many of the time consuming problems in scientific computation involve repeated evaluation of the same function on different argument sets. The argument sets which compose the problem data base must be structured in such a fashion that they can be distributed among 64 separate memories.

. . . .

Figure 1 illustrates that ILLIAC IV is designed to execute the same instruction streams simultaneously on 64 data sets that are stored in separate memories. Thus ILLIAC IV is well suited to evaluate the same function f on 64 sets of arguments, provided that the data base from which the argument sets are drawn can be structured in a certain fashion. The data base must be distributed among the PEs so that the argument sets required in the ith evaluation of the function f can be stored in the ith memory or in a memory that is "close" to the ith memory.

A primary requirement for many of the time consuming problems in scientific computation is repeated evaluation of the same functional form on different argument sets. Consider the following examples:

A. Matrix Multiply--The elements z_i of a column of the matrix Z = AX are produced by forming

$$z_i = \sum_{j=1}^{n} a_{ij}x_j \quad \text{where } x_j \text{ are the elements of the corres-}$$

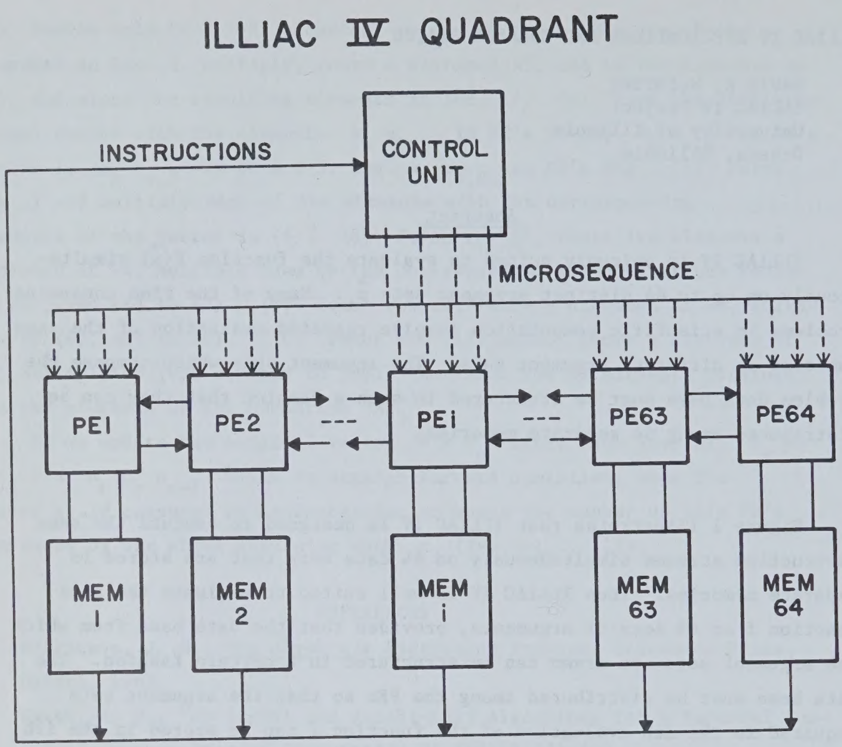
ponding column in X. This requires evaluation of the function

$$f(\vec{y}, \vec{x}) = \sum_{j=1}^{n} y_j x_j$$

where the argument

$(y_1, y_2, \ldots y_n)$ is repeatedly replaced by
$(a_{i1}, a_{i2}, \ldots a_{in})$ for $i = 1, 2, \ldots n$.

B. Solution of Simultaneous Linear Equations--Using the Gauss-Jordan method, the dominant computation is


ILLIAC IV QUADRANT
INSTRUCTIONS
CONTROL UNIT
MICROSEQUENCE
PE1
PE2

PEi

PE63
PE64
MEM 1
MEM 2
MEM i
MEM 63
MEM 64

FIGURE 1

$$a'_{ij} = a_{ij} - c_j a_{kj} \qquad \begin{array}{l} j = k,\ k+1, \ldots n \\ i = k,\ k+1, \ldots n \end{array}$$

or repeated evaluation of

$f(x, y, c) = x - cy$ where the three-tuple argument (x, y, c) is repeatedly replaced by (a_{ij}, a_{kj}, c_j).

C. Two-Dimensional Finite Difference Schemes--An explicit scheme for solving the "heat equation" $\frac{\partial u}{\partial t} = \nabla^2 u$ is

$$u_{ij}^{n+1} = u_{ij}^{n} + C\,(u_{i+1,j}^{n} + u_{i-1,j}^{n} + u_{i,j+1}^{n} + u_{i,j-1}^{n} - 4^{n}u_{ij}).$$

This requires that

$f(x,y,z,u,v,c) = x + c(y + z + u + v - 4x)$ be evaluated with suitable argument sets in which x takes on values of u_{ij} and y, z, u, v take on corresponding values of the "neighbors" of u_{ij}.

Note that if one sampled the instruction stream on a conventional machine during execution of these algorithms, the computation of f would dominate the calculation. For these illustrations, the data base can be stored among the distinct PE memories so that there is no difficulty in accessing the appropriate arguments and performing up to 64 simultaneous evaluations of f.

The two central considerations in programming ILLIAC IV are: (a) the exploitation of the simultaneous arithmetic capability and (b) the distribution of operands in the memories so that the required argument sets can be accessed without time consuming rearrangement of storage. Exploiting simultaneous arithmetic is generally straightforward. Often cursory inspection of the loops in a conventional program will reveal methods of utilizing the simultaneous arithmetic feature. There are cases where it is not straightforward, i.e., cases where, because of data dependent conditions or other considerations, it may be desirable to evaluate several different functional forms $f_1, f_2, \ldots f_k$ in different PEs at the same time. This can be accomplished by proceeding sequentially, evaluating

f_1 over PE set S_1

f_2 over PE set S_2

.

.

.

f_k over PE set S_k.

Alternately, remembering that all functions regardless of how dissimilar their forms, are composed of sequences of fetches, stores, and arithmetics, a "super" function

$$F = \bigcup_{i=1}^{k} f_i$$

can be constructed, which is the "union" of all instruction sequences defined by the f_i. Then the evaluation of F, interspersed with appropriate commands for enabling and disabling PEs, will yield the simultaneous evaluation of $f_1, f_2, \ldots f_k$.

As an elementary example, consider evaluating

$f_1(x,y) = x + y$ over PE set S_1

$f_2(x,y) = cx - y$ over PE set S_2.

Proceeding to evaluate f_1 and f_2 sequentially yields an instruction string:

$$\left.\begin{array}{l}\text{enable } S_1\\ \text{fetch}\\ \text{fetch}\\ \text{add}\end{array}\right\}\ \text{evaluation of } f_1$$

$$\left.\begin{array}{l}\text{enable } S_2\\ \text{fetch}\\ \text{fetch}\\ \text{multiply}\\ \text{fetch}\\ \text{subtract (add)}\\ \hline\end{array}\right\}\ \text{evaluation of } f_2$$

Total 5 fetches
2 adds
1 multiply
2 "enables"

But by "imbedding" the evaluation of f_1 in the evaluation of f_2 the following instruction stream is generated:

enable S_1 U S_2
fetch
enable S_2
fetch
multiply
enable S_1 U S_2
fetch
enable S_2
negate
enable S_1 U S_2
add

Total 3 fetches
1 add
1 multiply
1 negate
5 "enables"

Since the fetches and arithmetic instructions require about 7 clocks, and the negate and enables about 1, the latter method yields the fastest program. (See Ref. [1] for a more realistic example.)

Memory allocation is somewhat less straightforward. Generally if the data base for the problem lends itself to representation either in vector form or in matrix form there are schemes for storing the vectors or matrices among the 64 distinct memories which preserve the topology or connectedness of the elements in the data base. (Ref. [2, 3, 4, 5].) The function evaluation involved in example C requires that the PE memory which contains the element u_{ij} also have easy access to $u_{i,j+1}$, $u_{i,j-1}$, $u_{i+1,j}$, $u_{i-1,j}$. There are storage schemes which ensure that access to the "neighbors" of an element are always convenient. Furthermore there are schemes which allow simultaneous access of both rows and columns of a matrix.

The applicability of ILLIAC IV to large problems in scientific computation has been studied for a wide variety of problems. (Ref. [6].) As one might expect, problems have been found which are ideally suited for ILLIAC IV or which can be made ideally suited with very little modification to the algorithms. Problems have also been found which can be adapted for ILLIAC IV at some sacrifice in utilization of the PEs, and some problems have been found which defy modification to exploit the machine architecture. Table 1 gives some examples of types of calculations which fall in each of these categories.

Applicable with High PE Utilization

Finite difference calculations in 1, 2, and 3 dimensions
Matrix arithmetic
Quadrature (including fast Fourier transform)
Signal processing
Linear programming

Applicable with Moderate PE Utilization

Particle moving problems (non-linear Monte Carlo transport, etc.)
Non-symmetric eigenvalue calculation
Solution of linear equations and matrix inversion
Solution of non-linear equations
Polynomial root finding

Applicable with Low PE Utilization

Inversion of tri-diagonal matrices
Table look-up in large, unstructured tables

It is not uncommon to find that numerical techniques which are generally regarded as superior to others for conventional machines might not be superior for minimizing execution time on ILLIAC IV. For example, in the solution of linear equations Gaussian elimination and back substitution is generally regarded as superior to the Gauss-Jordan algorithm because there are fewer arithmetic operations involved. However, on ILLIAC IV, the execution time required to solve a set of linear equations employing the Gauss-Jordan algorithm is shorter than that required for Gaussian elimination and back substitution because, among other things, the back substitution portion of the algorithm is essentially a sequential process.

References

1. T. Yasui, "Programming the Tillotson Equation of State on ILLIAC IV", ILLIAC IV Document No. 189, University of Illinois, Urbana, Illinois, August 20, 1969.

2. R. Wilhelmson, "Matrix Operations on ILLIAC IV", ILLIAC IV Document No. 52, University of Illinois, Urbana, Illinois, March 1, 1967.

3. J. E. Stevens, "Matrix Multiplication Algorithms for ILLIAC IV", ILLIAC IV Document No. 231, University of Illinois, Urbana, Illinois, August 26, 1970.

4. V. Benokraitis, "Alternate Storage Methods for Two-Dimensional Hydrodynamics Calculations", ILLIAC IV Document No. 136, University of Illinois, Urbana, Illinois, May 27, 1968.

5. P. Kraška, "Array Storage Allocation", ILLIAC IV Document No. 186, University of Illinois, Urbana, Illinois, August 1, 1969.

6. "ILLIAC IV Research Document Abstracts", University of Illinois, Urbana, Illinois, July 1, 1971.

ILLIAC IV AND THE ARPA NETWORK

W. J. BOUKNIGHT
University of Illinois
Urbana, Illinois 61801

ABSTRACT

The ARPA Network, a highly sophisticated and wide ranging experiment in the remote access and sharing of computer resources, is briefly described and its current status discussed. Many researchers located about the country who will use ILLIAC IV in solving problems will do so via the network. The various systems, hardware and procedures they will use will be discussed.

INTRODUCTION

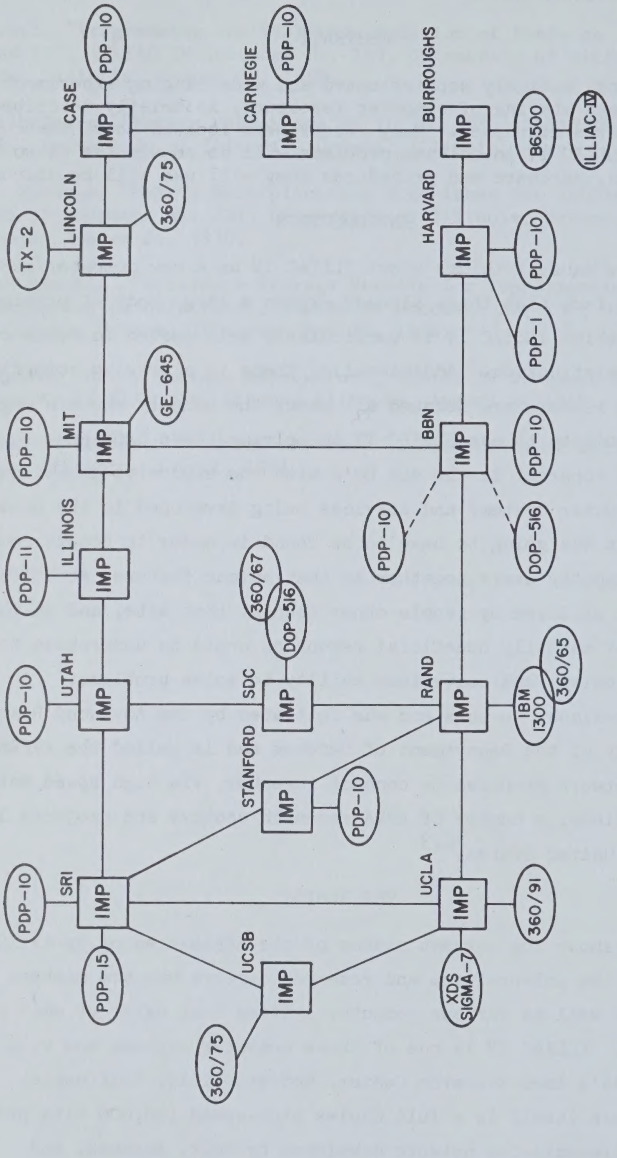
One of the unusual things about ILLIAC IV as a new computer architecture is the fact that there already exists a large body of problems to be solved for which ILLIAC IV is particularly well suited in terms of size, speed, and sophistication. Additionally, there is a growing community of scientists and researchers located all about the country who are eagerly awaiting the ability to use ILLIAC IV in solving these problems.

It became apparent in the mid 60's with the explosive growth and variety of computer systems and services being developed in the country that some means was going to have to be found in order to couple computer systems and computer users together so that unique features at different sites could be utilized by people other than at that site, and in this way, sharing of mutually beneficial resources could be undertaken to provide more economical and convenient ability to solve problems.[1] A prodigious experiment to this end was initiated by the Advanced Research Projects Agency of the Department of Defense and is called the ARPAnet. The ARPAnet network proposes to connect together, via high speed data transmission lines, a number of ARPA research centers and projects located all about the United States.[2,3]

THE DESIGN

Figure 1 shows the current status of the ARPAnet as of April 1971. Indicated are the universities and research centers who are members of the network as well as various computer systems that exist at each of those centers. ILLIAC IV is one of those computer systems and will be located at NASA's Ames Research Center, Moffett Field, California.

The network itself is a full duplex high-speed (50,000 bits per second) data transmission network developed by Bolt, Beranek, and


PDP-15
PDP-10
SRI
IMP
UCSB
IMP
360/75
UCLA
IMP
XDS SIGMA-7
360/91
STANFORD
IMP
PDP-10
PDP-10
UTAH
IMP
SDC
IMP
360/67
DOP-516
RAND
IMP
IBM 1300
360/65
PDP-11
ILLINOIS
IMP
PDP-10
MIT
IMP
GE-645
PDP-10
BBN
IMP
PDP-10
DOP-516
TX-2
LINCOLN
IMP
360/75
HARVARD
IMP
PDP-1
PDP-10
CASE
IMP
PDP-10
CARNEGIE
IMP
PDP-10
BURROUGHS
IMP
B6500
ILLIAC-IV

ARPANET, APRIL 1971

FIGURE I

Newman, Cambridge, Massachusetts.[*] The network itself is a store-and-forward message transmission network with the nodes of the network occupied by a sophisticated terminal called the interface message processor (IMP).[4] The heart of the IMP is a Honeywell DDP-516 computer which takes care of such tasks as error control, message routing, network "tuning" (adjusting its operation to maximize its efficiency), and statistics gathering. Much care has been devoted in the design and implementation of the network to insure an ultra-high level of reliability (currently no more than one single bit error per year should go undetected). By the store-and-forward mechanism, error control is implemented by retransmission of a message or section of message on which an error was detected. Error correcting codes are appended to each message to allow the detection of a wide class of errors and contribute to the reliability of transmission. All of the IMPs are passive devices in that they perform no other service besides message routing and control.

At any given node in the network, one or more HOST computers may be attached providing a service center or research project with access to the network.[**] The connection between a HOST and the IMP is made over a high speed interface at 100,000 bits per second, full duplex. Typical HOST computers in use around the network are the DEC PDP-10, IBM 360/75, IBM 360/91, IBM 360/67, and Burroughs B6500. While most of the HOST computers are associated with specific projects sponsored by ARPA, several locations are designated as service HOST sites. For instance, the 360/91 at the Campus Computing Network at UCLA is available on the network most of the time for use as a general computing service. Several other places, the University of California at Santa Barbara, Bolt Beranek and Newman in Cambridge, and MIT's Multics system also serve as HOST sites providing capabilities and services particular to their installations and computer systems.

---

[*] Full duplex means that the user has the ability to send and receive simultaneously--as opposed to half duplex where the user can only be sending or receiving at any instance in time.

[**] HOST refers to a computer system capable of supplying a full range of computer services to users on the network or at that site.

THE TOWER OF BABEL

The current conceptionalization of the ARPAnet presents, what some people call, "The Tower of Babel" problem in computer resource sharing. One can easily see that there are a wide variety of computer systems in geographically distant locations involved in many dissimilar projects, all speaking a variety of computer languages. While it was easy, relatively speaking, to bring up the hardware portion of the network, i.e., design and build IMP's, establish phone line connections for the 50,000 bit/second transmission lines, etc., much effort has been expended in trying to solve the "software" problem of the network; in other words, how does a human user talk to a remote HOST computer system, and also how does each HOST system talk to the other remote HOST systems.

For several years now, a collection of representatives from each site has been periodically meeting and corresponding with each other to design and implement a series of protocols (procedures) which lay down the ground rules of access and conversation between people and computers on the network. The most basic of these protocols is the IMP to HOST protocol.[5] This protocol contains the procedures by which each host talks to its IMP and data and routing information are exchanged.

The next level of protocol is that of HOST to HOST communication. This protocol contains the procedures and ground rules for information transfer between HOSTS and incorporates such features as connection procedures, space allocation, flow control, error control, etc. The HOST to HOST protocol was, in fact, the most difficult protocol to develop since it deals most closely with the specifics and idiosyncracies of the various computer hardware configurations and their respective operating systems.

Once the HOST to HOST protocol was formulated, a number of applications protocols have followed. These protocols in general allow the proper use of the HOST to HOST protocol by systems and applications programs desiring network access. The most important of these is the Initial Connection Protocol which allows the establishment of a full-duplex path of transmission between two systems or programs on the network. This protocol in general is masked from the user and is handled automatically by the network control programs in the HOST sites.

For transmission of data on the network, there is the Data Transfer Protocol which is concerned with the transmission of large blocks of data from point to point. Since most computer systems treat these large blocks of data as files and have various and elaborate file naming and access mechanisms implemented for the handling of these files, there is also a File Transfer Protocol which facilitates the handling of files between arbitrarily different systems on the network and allowing the mapping of their naming and access conventions. In addition to the variances in naming and access, there is the data formatting problem which exists in data transfers. The approach taken to solve this problem was to develop a data reconfiguration protocol and supporting systems whereby the user may specify a transformation function on his data string. He passes it through a reconfiguration service moving it from the sending site to the receiving site satisfying the format constraints at both sites. An example of this might be the conversion of 36-bit floating point format for the PDP-10 into 48-bit floating point format for the Burroughs B6500.

Most activities on the network for the individual user will be accomplished in a remote access manner to the various time sharing systems on the network. Since there is a wide variance in the types of terminals that exist on the network, a terminal network protocol (TELNET) was developed which attempts to mask all terminals into a network virtual terminal, pass the streams of data and control between that terminal and remote sites in both directions, then unmask the network virtual terminal at the receiving site into the form best desired by that site. For example, both Teletype terminals (33, 35, 37) and IBM 2741 terminals may have access to programs which expect one or the other of those terminals, both with equal facility.

The TELNET protocol handles the usual alphanumeric and hard copy terminals but does not handle the graphics terminals such as storage scopes or refresh graphics display terminals. To that end, a graphics network protocol is under development at this time. Results so far show, however, that a unified approach to the entire problem of graphics display and interaction is probably not feasible. Indeed, there may eventually be several protocols at various levels dealing with various classes of graphics devices and various modes of graphics display.

EXPANSION PLANS

With the initial success of the protocol efforts, the second phase of network development has been entered. During this phase, additional nodes will be added to the network. These nodes, however, differ widely from the IMP of the initial configuration. These new nodes provide for connection of terminal hardware directly to the network and therefore, are called

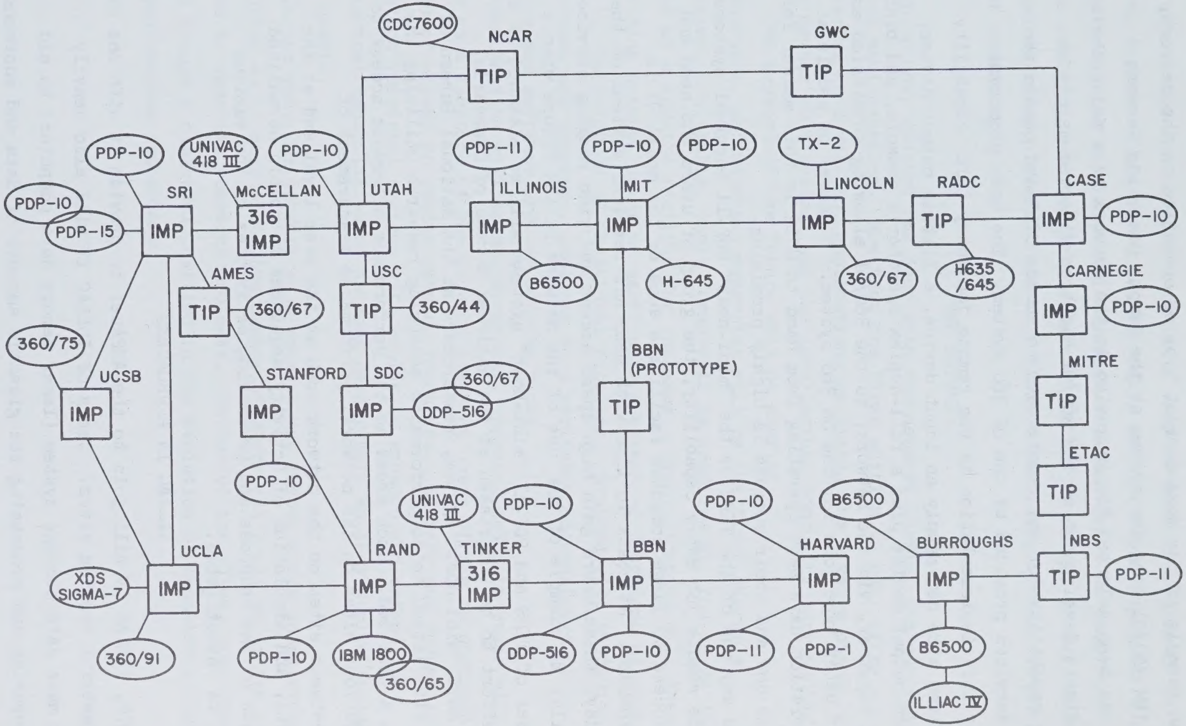
Terminal IMP's or TIP's. A TIP is a parasite node and provides no service capability to the network on its own. Users attached to the network via a TIP must derive all of the computational power and service from remote HOSTS on the network. Thus, the second phase of network development sees the introduction of user oriented groups to complement the present research and service organizations. Figure 2 depicts a projected arrangement of the network and its members as of April 1972.

In addition to the TIP, an effort has been underway for some time at the University of Illinois to develop a "mini HOST" computer system based on the configuration of a small mini-computer (DEC PDP-11) acting as a full capacity HOST (from the protocol standpoint) and attached to a standard IMP or a TIP. The PDP-11 based system (called the ARPA Network Terminal System or ANTS) serves essentially the same functions as a TIP. However, with the splitting of the network node and terminal functions into two separate computers, and providing additional capability in the PDP-11 in the form of a disk, a wider variety of peripherals, more memory, etc., a limited amount of on-site processing power is provided for the user to do housekeeping functions such as network accounting, on-site card deck-to-printer listings, data storage on magnetic tape and disk, etc., and to provide him with a higher level interface into the various protocol streams which will be directed to and from his site.

PROMISING RESULTS

A prominent example of network usage during its earliest implementation phase, when a subnet comprised of University of Utah, Stanford, UCLA and UCSB existed for a short time, occurred during the development of SRI's Intellect Augmentation research project. SRI had been using an XDS 940 computer to service their documentation aids and support system. During 1969-70, they were in the process of selecting and moving to a new central processing system to support that effort, a Digital Equipment Corporation PDP-10.

Prior to receiving their system, SRI had good contact with the University of Utah's graphics research project who already had a PDP-10. After the initial subnet implementation was completed, the SRI group had remote access to Utah's PDP-10 on which they began the conversion of their XDS 940 systems programs. The use of this subnet to connect their terminals and hard-copy devices (printers, etc.) to the Utah PDP-10 system, as if they were at the Utah site, enabled the systems programmers at SRI to complete their initial conversion project in less than half a year.


CDC7600
NCAR
TIP
GWC
TIP
PDP-10
UNIVAC 418 III
PDP-10
PDP-11
PDP-10
PDP-10
TX-2
PDP-10
PDP-15
SRI
IMP
McCELLAN
316 IMP
UTAH
IMP
ILLINOIS
IMP
MIT
IMP
LINCOLN
IMP
RADC
TIP
CASE
IMP
PDP-10
AMES
TIP
360/67
USC
TIP
360/44
B6500
H-645
360/67
H635 /645
CARNEGIE
IMP
PDP-10
360/75
UCSB
IMP
STANFORD
IMP
SDC
IMP
360/67
DDP-516
BBN (PROTOTYPE)
TIP
MITRE
TIP
ETAC
TIP
PDP-10
UNIVAC 418 III
PDP-10
PDP-10
B6500
UCLA
IMP
XDS SIGMA-7
RAND
IMP
TINKER
316 IMP
BBN
IMP
HARVARD
IMP
BURROUGHS
IMP
NBS
TIP
PDP-11
360/91
PDP-10
IBM 1800
360/65
DDP-516
PDP-10
PDP-11
PDP-1
B6500
ILLIAC IV

PROJECTED ARPANET, APRIL 1972

FIGURE 2

Activities during this time included remote text editing, compilations of systems programs, debug executions, and subsequent shipment of printer listings, debug results, etc., back to SRI for listing on their printer.

As an example of the more current type of operation on the network, the UCLA IBM 360/91 computer system at the Campus Computing Network division has been selected for inclusion in the network as a major service site. Software development on the 360/91 has been directed mainly at providing capability for any remote user to access UCLA and obtain the standard services provided by the OS 360 system. The most prominent feature of this implementation is the remote job entry (RJE) capability where a given site can supply an input device, a listing output device, and a punch output device plus a full-duplex operator's console, and bind these all together, via the network, to the 360/91 allowing submission and control of batch type jobs for the OS 360 system. A number of people at Rand Corporation have been operating from Rand to UCLA in this manner for some months now and their success is highly promising.

Since any one of the TIPS or the "mini-host" PDP-11 terminal systems can provide remote job entry capability, the group of users who need and can gain access to large computer facilities such as the model 91 is greatly expanded. At costs in line with many RJE terminal systems on the market today, these users gain high speed access and wide ranging services.

Finally, an example of the use of the network in the future when a large number of TIPS and PDP-11 "mini-host" systems will be attached is a research effort to be undertaken at the National Bureau of Standards. Using the ANTS "mini-host" system, researchers at the National Bureau of Standards site intend to do performance measuring research utilizing the network as a test bed. Such areas as the performance of remote access to and from various sites on the network, performance measurement of activity between sites on the network both at the user level and at the system level, and simulation of activity sequences to obtain normalized measurements of the response and processing activities of the various sites will be investigated.

ILLIAC IV APPROACHES

In 1972, ILLIAC IV will join to the ARPAnet to provide it with one of its most powerful service sites. Users of ILLIAC IV will also heavily access the mass data storage system (laser memory data computer) to aid them in preparing and processing the gigantic amounts of data and enormous problems which they will attempt to solve utilizing ILLIAC IV. Conversely, the future development of ILLIAC IV appears to benefit from the network in two ways. First, there are quite a number of people located all over the country who need direct access to ILLIAC IV for their particular applications and the network provides a very economical framework in which to dispense this service. Secondly, network users represent a large body of technical knowledge and experience in the craft of computer systems building, and the development of ILLIAC IV should be very much enriched by being placed within this flow of knowledge throughout the ARPAnet community.

THE FUTURE

Up to now, concern has been with the network itself, its concept and physical manifestation. Now that the first edition is out, efforts will be directed at developing applications utilizing the diverse capabilities of the member HOST systems and the intellectual community surrounding them. The keyword of the future will be sharing.

Hardware, software and expertise will be the first sharable quantities. Initially, their forms will be traditional, the large 360/91 at UCLA as a service facility for instance. Specialization will then set in yielding network usage of particular specialized functions at several sites in the solution of individual problems. Systems will subdivide to sub-systems and programs, evolving to a veritable smorgasboard of hardware and software problem solving tools.

Data sharing will become prominent. Specialized data bases will appear as a function of the growing needs of users of the network and their freedom of access to remote network resources. Library and documentation production retrieval systems will evolve to enhance and extend the intellectual capabilities of network researchers as they strive to master their quests.

In the end, the network as a concept may foster a new intellectual revolution as it multiplies the number of minds and tools which can be applied to specific research projects. As one of the first, the ARPAnet should occupy a classical position in the evolution of computer systems and applications technology.

BIBLIOGRAPHY

1. Marill, T. and Roberts, L. "Toward a Cooperative Network of Time-Shared Computers." Proceedings of FJCC, 1966.

2. Roberts, L. "The ARPA Network." Invitational Workshop on Networks of Computers Proceedings, National Security Administration, 1968.

3. Roberts, L. "Resource Sharing Computer Networks." IEEE International Conference, March, 1969.

4. Heart, F. E., Kahn, R. E., et al. "The Interface Message Processor for the ARPA Network." AFIPS Conference Proceedings, May, 1970.

5. Carr, S., Crocker, S., and Cerf, V. "HOST-HOST Communication Protocol in the ARPA Network." AFIPS Conference Proceedings, May, 1970.

ON OPTIMAL PAM MODULATION AND DEMODULATION

KAZUO MURANO AND JOHN B. THOMAS
Department of Electrical Engineering
Princeton University, Princeton, N.J. 08540

ABSTRACT

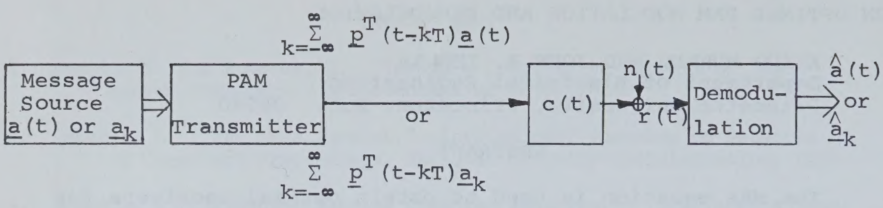
The MAP equation is used to obtain optimal receivers for both natural sampling (NS) and instantaneous sampling (IS) PAM systems. The optimal transmitter is also considered for vector IS PAM and it is shown that pure frequency domain multiplexing is the unique optimal solution.

I. INTRODUCTION

This paper considers two types of multi-input PAM systems: natural sampling (NS) PAM and instantaneous sampling (IS) PAM. Detailed description of these systems can be found in Rowe [1]. The problem is to obtain an optimal receiver and transmitter for the modulation systems. Tufts [2] has derived an optimal receiver using mean square error (MSE) as a criterion, while George [3] has shown that the same receiver is obtained when maximizing the signal-to-noise ratio. Berger and Tufts [4] have solved the problem of simultaneous optimization of transmitter and receiver for scalar IS PAM, and Dejon and Hänsler [5] have reported extensions to the vector modulation case.

In this paper, the optimum receiver is obtained using the MAP equation originally derived by Youla [6]. Although attention must be restricted to Gaussian noise and intelligence, this approach has the advantage that the structure of the receiver need not be pre-determined as has been the case in most previous investigations. The MAP equation itself implies a physical (though not necessarily realizable) structure and the principal difficulty lies in obtaining an equivalent realizable structure. The MAP equation has been successfully applied to the AM demodulation problem as in Thomas and Wong [7] and has been treated extensively in Van Trees [8]. In this paper, the realizable-with-delay optimum receivers are obtained for both NS and IS PAM, when the intelligence $\underline{a}(t)$ is an M-dimensional vector random process. Pulse shape optimization for vector IS PAM is also investigated. Dejon and Hänsler [5] have discussed a special case of this problem, where all sources were assumed to have identical statistics, and have found a class of optimum systems, frequency multiplexing and time multiplexing being special cases. In this paper, the source statistics are assumed to be distinct, and the result shows that, for the colored noise case, frequency multiplexing is the unique optimum system, an interesting contrast to the results of Dejon and Hänsler.

The systems to be investigated here are the two types of PAM as shown in Fig. 1. The message source $\underline{a}(t)$ is an M-dimensional Gaussian column vector random process, which is either naturally sampled (NS) by the M-dimensional column vector pulse train $\sum_{k=-\infty}^{\infty} \underline{p}(t-kT)$ or instantaneously sampled (IS) to


$\sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\underline{a}(t)$
Message Source
$\underline{a}(t)$ or $\underline{a}_k$
PAM Transmitter
or
$\sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\underline{a}_k$
c(t)
n(t)
r(t)
Demodulation
$\hat{\underline{a}}(t)$ or $\hat{\underline{a}}_k$

(Note: $c(t)$ is assumed to be $\delta(t)$ in the derivations given in this paper. Treatment of general $c(t)$ is straightforward [9].)

Fig. 1 - PAM Modulation System

obtain a random sequence $\{\underline{a}_k\}$ and then modulated by the pulse train. In both cases, the sampling interval is T seconds. The output of the transmitter is a multiplexed waveform $\sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\underline{a}(t)$ for NS PAM and $\sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\underline{a}_k$ for IS PAM. (The superscript T will be used to denote transpose.) The channel consists of a linear filter $c(t)$ and additive Gaussian noise $n(t)$. The problem is to obtain an optimal demodulator structure and optimal pulse shape $\underline{p}(t)$ for the system.

II. MAP RECEIVER FOR IS PAM

Assume that $\{\underline{a}_k\}_{k=-\infty}^{\infty}$ is a zero-mean, wide sense stationary and jointly Gaussian vector random sequence with covariance matrix $\underline{R}_a(k-\ell) = E\{\underline{a}_k \underline{a}^T_\ell\}$ and that the noise $n(t)$ is an independent zero-mean, continuous, and wide sense stationary Gaussian random process with covariance function $R_n(t-\tau) = E\{n(t)n(\tau)\}$. Then the MAP equation for the IS PAM system of Fig. 1 can be written as [8,9]

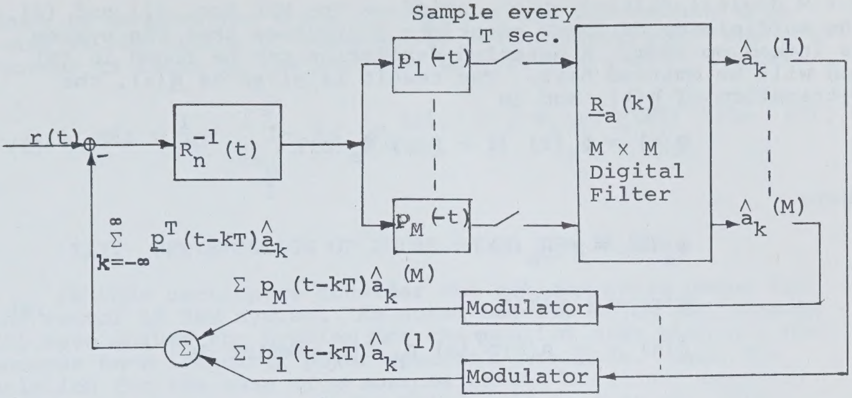
$$\hat{\underline{a}}_k = \sum_{\ell=-\infty}^{\infty} \underline{R}_a(k-\ell) \int_{-\infty}^{\infty} \underline{p}(z-\ell T)g(z)\,dz \tag{1}$$

and

$$r(t) - \sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\,\hat{\underline{a}}_k = \int_{-\infty}^{\infty} R_n(t-z)g(z)\,dz \tag{2}$$

It may be worthwhile to note here that, in general, the MAP equations are only necessary and not sufficient conditions for maximizing the a posteriori probability. But when the modulation is linear, as in this case, the a posteriori distribution of $\underline{a}_k$ given $r(t)$ becomes Gaussian and hence possesses a unique maximum. Thus the MAP equations are both necessary and sufficient. Furthermore, since the peak coincides with the conditional mean, the MAP receiver is also the least mean-squared error (MSE) receiver.

The structure of the receiver given explicitly by Eqs.(1) and (2) is shown in Fig. 2. Due to the feedback loops, this structure is inherently unrealizable, but a realizable-with-delay structure is obtained as follows. First, assume the system shown in Fig. 3 and find the expression for $\underline{h}(k)$ [an


Sample every T sec.
$p_1(-t)$
$p_M(-t)$
$r(t)$
$R_n^{-1}(t)$
$\underline{R}_a(k)$
M × M Digital Filter
$\hat{a}_k^{(1)}$
$\hat{a}_k^{(M)}$
$\sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\hat{\underline{a}}_k$
$\Sigma\, p_M(t-kT)\hat{a}_k^{(M)}$
$\Sigma\, p_1(t-kT)\hat{a}_k^{(1)}$
Modulator
Modulator
Σ

$p_i(t)$ is the i^{th} component of $\underline{p}(t)$

$\hat{a}_k^{(i)}$ " " " " " " " $\hat{\underline{a}}_k$

Fig. 2 - Unrealizable Demodulator

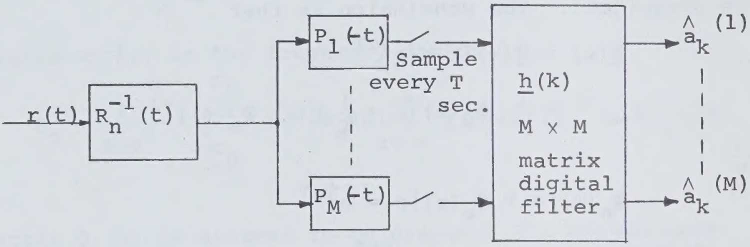
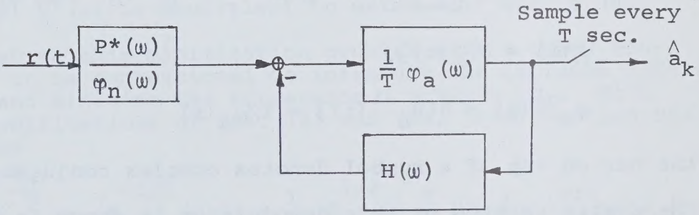

$r(t)$
$R_n^{-1}(t)$
$p_1(-t)$
$p_M(-t)$
Sample every T sec.
$\underline{h}(k)$
M × M matrix digital filter
$\hat{a}_k^{(1)}$
$\hat{a}_k^{(M)}$

Fig. 3 - Realizable-with-delay Demodulator


$r(t)$
$\frac{P^*(\omega)}{\varphi_n(\omega)}$
$\frac{1}{T}\varphi_a(\omega)$
Sample every T sec.
$\hat{a}_k$
$H(\omega)$

$$H(\omega) = \frac{1}{T}\sum_{k=-\infty}^{\infty} |P(\omega-k\omega_o)|^2 \varphi_n^{-1}(\omega-k\omega_o)$$

Fig. 4 - Scalar Case

M × M digital filter] which satisfies the MAP Eqs. (1) and (2). The sufficiency of these equations guarantees that the system is indeed optimum. A detailed derivation can be found in [9] and will be omitted here. The result is given as $\underline{H}(z)$, the z-transform of $\underline{h}(k)$, and is

$$\underline{H}(z) = \Phi_a(z)\ [I + \underline{F}(z)\ \Phi_a(z)]^{-1} \tag{3}$$

where

$$\begin{aligned} \Phi_a(z) &= \mathcal{Z}\{\underline{R}_a(k)\} \\ \underline{F}(z) &= \mathcal{Z}\{\underline{f}(k)\} \\ \underline{f}(k) &= \iint_{-\infty}^{\infty} \underline{P}(t)\underline{P}^T(s)\ R_n^{-1}(t-s+kT)dt\ ds \end{aligned} \tag{4}$$

Different realization of the receiver can be obtained if filtering and sampling are interchanged. Sampling is equivalent to translating all higher band components in the frequency domain to baseband. Let $\underline{H}(\omega)$ be the filter to be used when the order of sampling and filtering is reversed. Then obviously $\underline{H}(\omega)$ should be periodic with period $\omega_o = 2\pi/T$ and the baseband component should correspond to $\underline{H}(z)$ on the unit circle $z=\exp\{j\omega t\}$. The conclusion is that

$$\begin{aligned} \underline{H}(\omega) &= \underline{H}(z)|z = e^{j\omega T} \\ &= \frac{1}{T}\ \Phi_a(\omega)[I + \frac{1}{T}\underline{F}(\omega)\ \Phi_a(\omega)]^{-1} \end{aligned} \tag{5}$$

where

$$\begin{aligned} \Phi_a(\omega) &= T\ \Phi_a(z)|z = e^{j\omega T} \\ \underline{F}(\omega) &= \underline{F}(z)|z = e^{j\omega T} \\ &= \frac{1}{T}\sum_{k=-\infty}^{\infty} \underline{P}(\omega-k\omega_o)\overline{\underline{P}}^{\,T}(\omega-k\omega_o)\ \varphi_n^{-1}(\omega-k\omega_o) \\ \underline{P}(\omega) &= \mathfrak{F}\{\underline{P}(t)\}, \\ \varphi_n^{-1}(\omega) &= \mathfrak{F}\{R_n^{-1}(t)\} = 1/\varphi_n(\omega) \end{aligned} \tag{6}$$

Here the bar on top of a symbol denotes complex conjugate.

The scalar version of this demodulator is shown in Fig. 4. This is exactly the same receiver obtained by Tufts [2] and George [3]. The performance of the system with the optimum demodulator will be measured by the residual MSE.

$$\begin{aligned} \text{MSE} &= E\{(\hat{\underline{a}}_k - \underline{a}_k)^T\ (\hat{\underline{a}}_k - \underline{a}_k)\} \\ &= \frac{1}{2\pi j}\oint_c \text{tr}\ \left\{\Phi_a(z)[I + \Phi_a(z)\underline{F}(z)]^{-1}\right\}\frac{dz}{z} \end{aligned} \tag{7}$$

where tr denotes the trace of a matrix and the contour C is taken in the region of convergence of the integrand. When the contour is taken on the unit circle, letting $z=e^{j\omega t}$, we have

$$\mathrm{MSE} = \frac{1}{2\pi}\int_{-\frac{\omega_o}{2}}^{\frac{\omega_o}{2}} \mathrm{tr}\left\{\Phi_a(\omega)\,[I + \frac{1}{T}\,\Phi_a(\omega)\,\underline{F}(\omega)]^{-1}\right\}d\omega \qquad (8)$$

III. OPTIMIZATION OF PULSE SHAPE FOR IS PAM

In this section we consider the optimum pulse shape for the vector IS PAM system. As noted before, Dejon and Hänsler [5] have solved the problem for the special case when all the sources have identical power spectral densities. Here the solution for the case of M sources having different mutually correlated power spectral densities will be derived. The criterion of optimization is the MSE given by Eq. (8). The power constraint E_o will be imposed on the modulated pulse train as

$$E_o = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} E\left\{[\sum_{k=-\infty}^{\infty} \underline{p}^T(t-kT)\,\underline{a}_k]^2\right\}dt$$

or, equivalently, in the frequency domain, as

$$E_o = \frac{1}{2\pi T^2}\int_{-\frac{\omega_o}{2}}^{\frac{\omega_o}{2}} \mathrm{tr}\left\{\Phi_a(\omega)\sum_{k=-\infty}^{\infty} \underline{P}(\omega-k\omega_o)\,\overline{\underline{P}}^{\,T}(\omega-k\omega_o)\right\}d\omega \qquad (9)$$

The matrix $\Phi_a(\omega)$ is assumed to be diagonal for convenience. There is no loss of generality in this assumption since Φ_a is Hermitian and there always exists a unitary matrix $U(\omega)$ such that $U\,\Phi_a\,\overline{U}^T$ is diagonal. It is clear from Eqs. (8) and (9) that using $U(\omega)\,P(\omega)$ with power spectral density matrix $U(\omega)\,\Phi_a(\omega)\,\overline{U}^T(\omega)$ is equivalent to using $P(\omega)$ with $\Phi_a(\omega)$.

To solve this minimization problem with a power constraint, it is conventional to introduce the Lagrange multiplier λ and minimize the expression $J = \mathrm{MSE} + \lambda E_o$. With slight modifications of Eqs. (8) and (9), this function can be written as

$$J = \frac{1}{2\pi}\int_{-\frac{\omega_o}{2}}^{\frac{\omega_o}{2}} \mathrm{tr}\left\{\Phi_a^{1/2}(\omega)\,[I + \frac{1}{T^2}\,\Phi_a^{1/2}(\omega)\sum_{k=-\infty}^{\infty}\underline{P}(\omega-k\omega_o)\overline{\underline{P}}^T(\omega-k\omega_o)\right.$$

$$\cdot\,\varphi_n^{-1}(\omega-k\omega_o)\,\Phi_a^{1/2}(\omega)]^{-1}\,\Phi_a^{1/2}(\omega)$$

$$\left.+\,\frac{\lambda}{T^2}\,\Phi_a^{1/2}(\omega)\sum_{k=-\infty}^{\infty}\underline{P}(\omega-k\omega_o)\overline{\underline{P}}^T(\omega-k\omega_o)\,\Phi_a^{1/2}(\omega)\right\}d\omega \qquad (10)$$

Using the techniques of the calculus of variation, we let $\underline{P}(\omega) = \underline{P}_o(\omega) + \epsilon\,\Delta\,\underline{P}_o(\omega)$, where $\Delta\,\underline{P}_o(\omega)$ is an arbitrary variation. Also let $\partial J/\partial\epsilon\big|_{\epsilon=o} = 0$. Then the result is, for all $\omega \in [-\infty,\infty]$,

$$\left[\Phi_a^{1/2}(\omega)G^{-1}(\omega)\Phi_a(\omega)G^{-1}(\omega)\Phi_a^{1/2}(\omega)\right]\underline{P}_o(\omega) = \lambda\,\varphi_n(\omega)\Phi_a(\omega)\underline{P}_o(\omega)$$

where (11)

$$G(\omega) = \left[I + \frac{1}{T^2}\Phi_a^{1/2}(\omega)\sum_{k=-\infty}^{\infty}\underline{P}_o(\omega-k\omega_o)\underline{P}_o^T(\omega-k\omega_o)\varphi_n^{-1}(\omega-k\omega_o)\Phi_a^{1/2}(\omega)\right]$$

Equation (11) can be rewritten as

$$\left[G^{-1}(\omega)\Phi_a(\omega)G^{-1}(\omega)\right]\Phi_a^{1/2}(\omega)\underline{P}_o(\omega) = \lambda\,\varphi_n(\omega)\Phi_a^{1/2}(\omega)\underline{P}_o(\omega) \tag{12}$$

At this point the noise will be assumed to be colored in the following sense: For almost all $\omega \in [-\omega_o/2, \omega_o/2]$, $\varphi_n(\omega-k_1\omega_o) \neq \varphi_n(\omega-k_2\omega_o)$ for any pair of integers k_1, k_2, such that $k_1 \neq k_2$.

Dejon and Hänsler [5] have proved that the total measure where $\underline{P}_o(\omega)$ is non-zero is bounded by $M\omega_o$. Using this fact, the summation in Eqs. (8) and (9) can be reduced to a finite sum of M terms and

$$\mathrm{MSE} = \frac{1}{2\pi}\int_{\frac{-\omega_o}{2}}^{\frac{\omega_o}{2}} \mathrm{tr}\left\{\Phi_a(\omega)\left[I + \frac{1}{T^2}\Phi_a(\omega)\sum_{i=1}^{M}\underline{P}(\omega-k_i\omega_o)\underline{P}^T(\omega-k_i\omega_o)\varphi_n^{-1}(\omega-k_i\omega_o)\right]^{-1}\right\}d\omega \tag{13}$$

and

$$E_o = \frac{1}{2\pi T^2}\int_{\frac{-\omega_o}{2}}^{\frac{\omega_o}{2}} \mathrm{tr}\left\{\Phi_a(\omega)\sum_{i=1}^{M}\underline{P}(\omega-k_i\omega_o)\underline{P}^T(\omega-k_i\omega_o)\right\}d\omega \tag{14}$$

Now define two new M × M matrices $Q(\omega)$ and $N(\omega)$ as

$$Q(\omega) = \left[\underline{P}(\omega-k_1\omega_o), \underline{P}(\omega-k_2\omega_o), \ldots, \underline{P}(\omega-k_M\omega_o)\right] \tag{15}$$

and

$$N(\omega) = \begin{bmatrix} \varphi_n(\omega-k_1\omega_o) & & & 0 \\ & \varphi_n(\omega-k_2\omega_o) & & \\ & & \ddots & \\ 0 & & & \varphi_n(\omega-k_M\omega_o) \end{bmatrix} \tag{16}$$

Then Eqs. (13) and (14) can be rewritten as,

$$\text{MSE} = \frac{1}{2\pi} \int_{-\omega_o/2}^{\omega_o/2} \text{tr} \left\{\Phi_a(\omega)\left[I + \frac{1}{T^2}\Phi_a(\omega)Q(\omega)N^{-1}(\omega)\overline{Q}^T(\omega)\right]^{-1}\right\}d\omega \tag{17}$$

and

$$E_o = \frac{1}{2\pi T^2}\int_{-\omega_o/2}^{\omega_o/2} \text{tr}\left\{\Phi_a(\omega)Q(\omega)\overline{Q}^T(\omega)\right\}d\omega \tag{18}$$

The problem is to minimize Eq. (17) under the constraint of Eq. (18). A similar situation appears in a previous work by Pilc [10], where he obtained optimum pre-and-post matrix filtering for multichannel transmission of a discrete memoryless Gaussian vector source. He proved that the optimum pre-and-post filters must be diagonal. Some of his ideas can be applied to the present problem. In the following, it will be proved that pure frequency domain multiplexing is the method which achieves the minimum MSE.

LEMMA Let A be an M-dimensional non-negative Hermitian matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq 0$. Also let B be an M × M non-negative diagonal matrix with diagonal elements $b_1 \geq b_2 \geq \ldots \geq b_M \geq 0$. Then

$$\text{tr}\ \{B\Lambda\} = \min_{U:\ \text{unitary}} \text{tr}\ \{B\overline{U}^T A U\}$$

where

$$\Lambda = \begin{bmatrix} \lambda_M & & & 0 \\ & \lambda_{M-1} & & \\ & & \ddots & \\ 0 & & & \lambda_1 \end{bmatrix}$$

As shown by Eqs. (17) and (18), interchanging the diagonal elements of N or Φ_a does not affect the result, since interchanging the corresponding rows or columns, respectively, of Q will give an equivalent system. Hence it may be assumed without loss of generality that the components of N and Φ_a are ordered in the following fashion:

$$\varphi_a^{(1)} \geq \varphi_a^{(2)} \geq \ldots\ldots\ldots \geq \varphi_a^{(M)}$$

$$\varphi_{n_1} \leq \varphi_{n_2} \leq \ldots\ldots\ldots \leq \varphi_{n_M}$$

where $\varphi_a^{(i)}$ denotes the power spectral density of the i^{th} source and $\varphi_{n_i} = \varphi_n(\omega - k_i\omega_o)$.

THEOREM The minimum MSE given by Eq. (17) under the power constraint of Eq. (18) can be achieved by a diagonal matrix Q_o.

PROOF First, slight modification of Eqs. (17) and (18) appears useful. We write

$$\text{MSE} = \frac{1}{2\pi}\int_{-\omega_o/2}^{\omega_o/2} \text{tr}\left\{\Phi_a(\omega)\left[I + \frac{1}{T^2}\Phi_a^{1/2}(\omega)Q(\omega)N^{-1}(\omega)\bar{Q}^T(\omega) \cdot \Phi_a^{1/2}(\omega)\right]^{-1}\right\}d\omega \qquad (19)$$

and

$$E_o = \frac{1}{2\pi T^2}\int_{-\omega_o/2}^{\omega_o/2} \text{tr}\left\{\bar{Q}^T(\omega)\Phi_a Q(\omega)\right\}d\omega \qquad (20)$$

Thus the two matrices $\left[I + T^{-2}\Phi_a^{1/2}QN^{-1}\bar{Q}^T\Phi_a^{1/2}\right]^{-1}$ and $\bar{Q}^T\Phi_a Q$ become non-negative Hermitian matrices. Now assume that Q is any arbitrary pulse matrix. Then, for each frequency $\omega \in [-\omega_o/2, \omega_o/2]$, the matrix $[I + T^{-2}\Phi_a^{1/2}QN^{-1}\bar{Q}^T\Phi_a^{1/2}]$ has M eigenvalues $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_M$. Since $\left[\Phi_a^{1/2}QN^{-1}\bar{Q}^T\Phi_a^{1/2}\right]$ is non-negative, $\eta_i \geq 1$ for $1 \leq i \leq M$. The main idea of the proof is that it is possible to reduce the MSE and power at the same time by using a diagonal matrix Q_o, while keeping the eigenvalues η_i fixed.

(1) Reduction of MSE
By the Lemma, the integrand of MSE (19) is a minimum when

$$\left[I + T^{-2}\Phi_a^{1/2}QN^{-1}\bar{Q}^T\Phi_a^{1/2}\right]^{-1} = \begin{bmatrix} \eta_1^{-1} & & & 0 \\ & \eta_2^{-1} & & \\ & & \ddots & \\ 0 & & & \eta_M^{-1} \end{bmatrix} \qquad (21)$$

(2) Reduction of Power
Since inverting the order of multiplication of two square matrices does not affect the eigenvalues of the product matrix, the eigenvalues of $\left[T^{-2}N^{-1}\bar{Q}^T\Phi_a Q\right]$ are the same as those of $\left[T^{-2}\Phi_a^{1/2}QN^{-1}\bar{Q}^T\Phi_a^{1/2}\right]$ that is $(\eta_i - 1)$, $1 \leq i \leq M$. Hence by Lemma, the integrand of the power expression of Eq. (20) is a minimum when

$$\frac{1}{T^2}N^{-1}\bar{Q}^T\Phi_a Q = \begin{bmatrix} \eta_1 - 1 & & & 0 \\ & \eta_2 - 1 & & \\ & & \ddots & \\ 0 & & & \eta_M - 1 \end{bmatrix} \qquad (22)$$

Now it is easy to prove that

$$Q_o = \begin{bmatrix} q_1 & & & 0 \\ & q_2 & & \\ & & \ddots & \\ 0 & & & q_M \end{bmatrix} \qquad (23)$$

where $q_i = \left(\varphi_{n_i}(\eta_i-1)T/\varphi_a^{(i)}\right)^{1/2}$ realizes Eqs. (21) and (22).

The arguments given thus far can be summarized as: While the total bandwidth of $\underline{P}_o(\omega)$ is at most $M\omega_o$, the total bandwidth of each component of $\underline{P}_o(\omega)$ is at most ω_o and the carrier sets for each of the components are mutually disjoint. Furthermore, the source with the largest value of $\varphi_a^{(i_1)}(\omega)$ is combined with the smallest noise $\varphi_n(\omega-k_{j_1}\omega_o)$ in which case the i_1^{th} component of $\underline{P}_o(\omega-k_{j_1}\omega_o)$ takes a non-zero value. Similarly, the second largest signal $\varphi_a^{(i_2)}(\omega)$ is combined with the second smallest noise $\varphi_n(\omega-k_{j_2}\omega_o)$, etc.

Now it is easy to show that the carrier set for $\underline{P}_o(\omega)$ can be given as follows: for each $\omega \in [-\omega_o/2, \omega_o/2]$ pick M integers $\{k_i\}_{i=1}^{M}$ such that $\varphi_n(\omega-k_1\omega_o) \le \varphi_n(\omega-k_2\omega_o) \le \ldots \le \varphi_n(\omega-k_M\omega_o) \le \varphi_n(\omega-\ell\omega_o)$ for all $\ell \notin \{k_i\}_{i=1}^{M}$. At this point the problem has degenerated to N parallel scalar problems and the solution can be given easily by referring to earlier results [4]. The solution is given as an algorithm in eight steps.

Step 1. For each $\omega \in [-\omega_o/2, \omega_o/2]$ pick the M smallest values of $\{\varphi_n(\omega-k\omega_o)\}_{k=-\infty}^{\infty}$. These are denoted by $\varphi_n(\omega-k_1\omega_o)$, $\varphi_n(\omega-k_2\omega_o), \ldots, \varphi_n(\omega-k_M\omega_o)$, in ascending order.

Step 2. Pick the largest of $\{\varphi_a^{(j)}\}_{j=1}^{M}$ and call it $\varphi_a^{(i_1)}(\omega)$. Now the vector $\underline{P}_o(\omega-k_1\omega_o)$ is assigned to the i_1^{th} source. Similarly, if $\varphi_a^{(i_n)}$ is the n^{th} largest of $\{\varphi_a^{(j)}\}_{j=1}^{M}$ then $\underline{P}_o(\omega-k_n\omega_o)$ is assigned to the i_n^{th} source.

Step 3. Fix a constant λ.

Step 4. Calculate the total energy by the formula

$$E_o^{1} = \frac{1}{2\pi}\int_{-\omega_o/2}^{\omega_o/2}\sum_{j=1}^{M} \text{Max}\left\{\left(\varphi_n(\omega-k_j\omega_o)\varphi_a^{(i_j)}(\omega)/\lambda\right)^{1/2} - \varphi_n(\omega-k_j\omega_o), 0\right\} d\omega \qquad (24)$$

Step 5. Adjust the value of λ according to the value of E_o^{1}. If E_o^{1} is greater than the original power constraint E_o, then λ should be made larger and vice versa.

Step 6. Repeat Steps 3 to 5 until $E_o^1 = E_o$.

Step 7. For each $\omega \in [-\omega_o/2, \omega_o/2]$ determine the absolute value of the i_j^{th} $(1 \le j \le M)$ component of vector $\underline{P}_o(\omega - k_j\omega_o)$ by

$$|P_{ij}(\omega-k_j\omega_o)|^2 = \text{Max}\left\{\left(\varphi_n(\omega-k_j\omega_o)/\lambda\, \varphi_a^{(i_j)}(\omega)\right)^{1/2} - \varphi_n(\omega-k_j\omega_o)/\varphi_a^{(i_j)}(\omega),\ 0\right\} \tag{25}$$

The phase of $\underline{P}_o(\omega-k_j\omega_o)$ is arbitrary.

Step 8. Calculate the MSE by

$$\text{MSE} = \frac{1}{2\pi}\int_{-\omega_o/2}^{\omega_o/2} \sum_{j=1}^{M} \text{Min}\left\{\left(\lambda\varphi_n(\omega-k_j\omega_o)\varphi_a^{(i_j)}(\omega)\right)^{1/2}, \varphi_a^{(i_j)}(\omega)\right\}d\omega \tag{26}$$

The derivation of formulas (24) - (26) is analogous to that given in [4] and will be omitted here.

The theorem states only that the diagonal matrix Q_o achieves the minimum MSE. No guarantee is provided that there is no other pulse matrix Q_o' which also attains this minimum. This point can be clarified [9] and the conclusion is that when the power spectral densities of the M sources are all distinct and the noise is colored, pure frequency domain multiplexing system is the unique optimum solution. In the white noise case, a whole class of systems is found to be optimum even for distinct source power spectral densities.

IV. MAP RECEIVER FOR NS PAM

Although the IS PAM systems were treated extensively in the previous sections, NS PAM will be only briefly investigated here. The MAP equations will be stated and the realizable-with-delay demodulators will be given. The resulting MSE will also be calculated, but no work will be done on the optimization of the pulse shape in this case.

The MAP equations can be obtained easily [8] as

$$\underline{\hat{a}}(t) = \int_0^T \underline{R}_a(t,z)\underline{S}(z)g(z)dz \tag{27}$$

and

$$r(t) - \underline{S}^T(t)\underline{\hat{a}}(t) = \int_0^T R_n(t,z)g(z)dz \tag{28}$$

where

$$\underline{S}(z) = \sum_{k=-\infty}^{\infty} \underline{p}(z-kT)$$

As before these equations do not give a realizable demodulator explicitly. When $\underline{a}(t)$ is bandlimited, following the same procedure as previously, assuming wide-sense stationarity, and allowing the observation time $[0,T]$ to be $[-\infty,\infty]$, the matrix receiver $\underline{H}(\omega)$ analogous to the matrix receiver for vector IS PAM case is obtained as in Fig. 5:

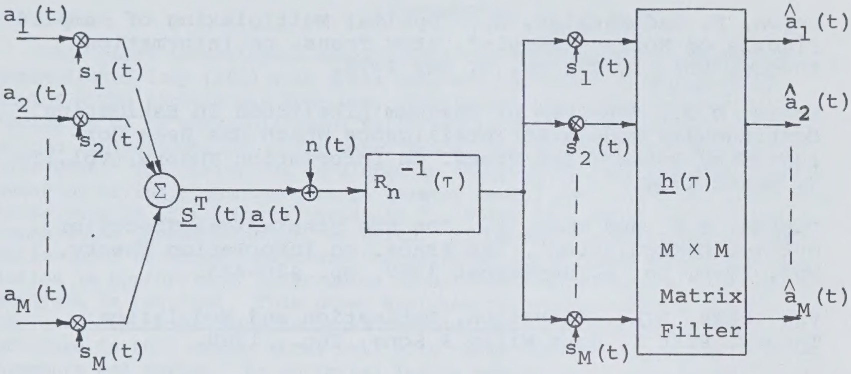


Fig. 5 - MAP Receiver for NS PAM

$$\underline{H}(\omega) = \Phi_a(\omega)\left[I + \sum_{k=-\infty}^{\infty} \underline{c}_{-k}\, \underline{c}_k^T\, \varphi_n^{-1}(\omega - k\omega_o)\Phi_a(\omega)\right]^{-1} \tag{29}$$

where $\Phi_a(\omega) = \mathfrak{F}\{\underline{R}_a(\tau)\}$ and $\underline{c}_k$ is the k^{th} Fourier coefficient of $\underline{S}(t)$. The resulting MSE can be calculated to be

$$MSE = \frac{1}{2\pi}\int_{-\omega_o/2}^{\omega_o/2} \mathrm{tr}\left\{\Phi_a(\omega)\left[I+\Phi_a(\omega)\sum_{k=-\infty}^{\infty}\underline{c}_{-k}\,\underline{c}_k^T\,\varphi_n^{-1}(\omega-k\omega_o)\right]^{-1}\right\}d\omega \tag{30}$$

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the National Science Foundation under Grants GK-1439 and GK-24187 and of the U.S. Army Research Office under Contract DAHC04-69-0012.

REFERENCES

1. Rowe, H., Signals and Noise in Communication Systems , Van Nostrand Company, 1965.

2. Tufts, D.W., "Nyquist's Problem - The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation", Proc. of IEEE, Vol. 53, No. 3, March 1965.

3. George, D.A., "Matched Filters for Interfering Signals", IEEE Trans. on Information Theory, Vol. IT-11, No. 1,

January 1965.

4. Berger, T. and Tufts, D., "Optimum Pulse Amplitude Modulation, Part I = Transmitter - Receiver Design and Bounds from Information Theory", IEEE Trans. on Information Theory, Vol. IT-13, No. 2, April 1967.

5. Dejon, F. and Hänsler, E., "Optimal Multiplexing of Sampled Signals on Noisy Channels", IEEE Trans. on Information Theory, Vol. IT-17, No. 3, May 1971.

6. Youla, D.C., "The Use of Maximum Likelihood in Estimating Continuously Modulated Intelligence Which Has Been Corrupted by Noise", IRE Trans. on Information Theory, Vol.IT-3, March 1954.

7. Thomas, J.B. and Wong, E., "On the Statistical Theory of Optimum Demodulation", IRE Trans. on Information Theory, Vol. IT-6, No. 4, September 1960, pp. 420-425.

8. Van Trees, H.L., Detection, Estimation and Modulation Theory, Part I, John Wiley & Sons, Inc., 1968.

9. Murano, K., Ph.D. Dissertation in preparation, Department of Electrical Engineering, Princeton University.

10. Pilc, R.J., "The Optimum Linear Modulator for a Gaussian Source Used with a Gaussian Channel", Bell Systems Tech. Jour., Vol. 48, Nov. 1969.

THRESHOLD PERFORMANCE OF A PHASE-LOCKED LOOP WITH PHASE FEEDBACK

JACOB KLAPPER
Department of Electrical Engineering
Newark College of Engineering
Newark, New Jersey

G. M. VEIGA
Western Electric Co.
Newark, New Jersey

INTRODUCTION

This paper investigates analytically the FM threshold performance of a phase-locked loop (PLL) with phase feedback described by Acampora and Newton.[1] The circuit has some very interesting and useful properties, including an FM threshold level which may be considerably lower than that of the conventional PLL. The experimental results were also duplicated elsewhere.[2] The device was developed in the pursuit of a PLL with a phase detector having a monotonic range greater than ± 90°, which has been shown to be capable of a lowered threshold.[3] It was thus named the Extended Range Phase Locked Detector, or ERPLD. It became however apparent that while under no noise conditions the extended range phase detector formulation is appropriate, under noise interference conditions a more extensive analysis is required. This paper analyzes the system without resorting to the extended range phase detector formulation. The analytical results obtained in this manner agree well with those obtained experimentally by Acampora and Newton. An optimized design presented in this paper offers some further threshold extension capability.

A block diagram of the Acampora and Newton circuit is shown in Figure 1. The block which is additional to the conventional second-order PLL is the phase modulator of sensitivity K_3. The analysis we shall present follows closely that used for the conventional PLL.[4] Beginning with the linear equivalent circuit (Fig. 2) which holds for small dynamic phase error conditions the various response functions are derived which hold above threshold. Threshold is then defined as the condition that the mean-square phase error at the phase detector output $\overline{\phi_e^2}$ has a certain value ν . Experience with conventional phase locked loops indicates that $\nu = 0.2$ to 0.25 is an indication of the point where the output signal-to-noise ratio is 1dB below that predicted from the above threshold formulation. The optimum design parameters are then obtained by minimizing the required carrier power to keep the system above threshold under fixed noise interference conditions. The loop is of second order, with $a \gg b$, and $K \gg b$ (K= basic PLL loop gain = $\delta K_1 K_2$) consistent with current practice in FM PLL demodulation.

CLOSED-LOOP RESPONSE

The closed-loop response from the input to the loop to the output of the VCO may be described in terms of the normalized second-order response (See Apprendix A).

$$H(s) = \frac{\Phi_r(s)}{\Phi_i(s)} = \frac{\frac{s}{a} + 1}{\left(\frac{s}{\omega_n}\right)^2 + 2\xi\left(\frac{s}{\omega_n}\right) + 1} \qquad (1)$$

Where $$\omega_n^2 = \frac{Kb}{\alpha_1} \quad , \quad \xi = \frac{1}{2}\left(\frac{\alpha_1 b}{K}\right)^{1/2} + \frac{1}{2a}\left(\frac{Kb}{\alpha_1}\right)^{1/2} \tag{2}$$

$$\alpha_1 = 1 + \alpha_2 \quad , \quad \alpha_2 = K_3 \delta \quad , \quad K = \delta K_1 K_2$$

The parameters K_1, K_2, K_3, a, b, and δ are defined in figures 1 and 2.

NOISE BANDWIDTH

The noise bandwidth of the loop is given by

$$B_n = \int_0^\infty \left|H(j\omega)\right|^2 df^{*} = \frac{Kb}{4\alpha_1 a}\left[\frac{\frac{Kb}{\alpha_1 a} + a}{\frac{Kb}{\alpha_1 a} + b}\right]$$

$$\approx \frac{1}{4}\left[\frac{Kb}{\alpha_1 a} + a\right] \quad \text{for} \quad \frac{K}{\alpha_1} \gg a \tag{3}$$

MEAN-SQUARE PHASE ERROR

The mean-square phase error at the phase detector output due to both signal and noise is given by

$$\overline{\phi_e^2} = \overline{(\phi_i - \phi_{ri})^2} = \frac{Y}{(Kb)^2} + B\frac{\eta}{A^2} \tag{4}$$

where $$Y = \int_0^{f_b} \omega^4 W_{\phi_i}\, df \tag{5}$$

$$\phi_{ri} = \phi_r + K_3 e \tag{6}$$

W_{ϕ_i} = single sided power spectral density of the input signal phase modulation,
η = single sided input noise power spectral density (assumed flat),
A = amplitude of input carrier,
f_b = top baseband frequency,
and B is given by

$$B = \frac{a}{2\alpha_1^2} + \frac{Kb}{2a\alpha_1^3}(\alpha_1 + \alpha_2) + \left(\frac{\alpha_2}{\alpha_1}\right)^2 B_{IF} \tag{7}$$

where BIF is the predetection noise bandwidth.
Equation 4 is derived in Appendix A.

---

* The solutions to integrals of this type are tabulated;
See e. g. Ref. 4, Ch 2.

THRESHOLD CNR

The threshold carrier-to-noise ratio, with the noise measured in twice basebandwidth is given by

$$\left[CNR_{2f_b}\right]_{TH} = \frac{B}{4f_b\left[\nu - \frac{Y}{(Kb)^2}\right]} \tag{8}$$

Equation 8 is derived in Appendix B.
Note that unlike the PLL, the threshold CNR of this loop is a function of the predetection filtering (through B). This fact is not recognizable from the extended range phase detector formulation.

OPTIMUM DESIGN

The optimum loop parameters for lowest threshold CNR are shown in Appendix C to be given by the equations

$$a = \left[\frac{Kb}{\alpha_1}\left(1 + 2\alpha_2\right)\right]^{1/2} \tag{9}$$

$$(Kb)^{5/2} = \frac{5Y}{\nu}(Kb)^{1/2} + \frac{4Y}{\nu}\alpha_2^2 B_{IF}\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)^{1/2} \tag{10}$$

A digital computer solution of the above equations indicates that $\alpha_2=1$, as used by Acampora and Newton[1], is a reasonably optimum value for a large range of system parameters. Also, a threshold CNR (with the noise measured in the predetection bandwidth) of 1.8 dB is obtained for the modulation and loop parameters used by Acampora and Newton, which agrees with their experimental results. (ν is taken as 0.25 as in the conventional PLL). For their system parameters, the above formula indicate that the PLL with phase feedback is better than the conventional PLL only for $B_{IF}/a < 2.75$, and is worse for larger ratios. Using the optimum parameters given above, a further threshold extension of 1.2 dB beyond that achieved by Acampora and Newton is indicated for their test-tone modulation, resulting for that case in a threshold CNR which is lower by 4.2 dB than that in the conventional optimum PLL.

CONCLUSION

An analysis was given for the PLL with phase feedback similar to that available for the conventional PLL. Its performance equations and an optimum design are derived. It is shown that the loop is capable of considerably better threshold performance than the conventional PLL, in agreement with the results of Acampora and Newton. However, with the assumed wideband phase feedback the performance is limited by the predetection noise bandwidth; such is not the case in the conventional PLL.

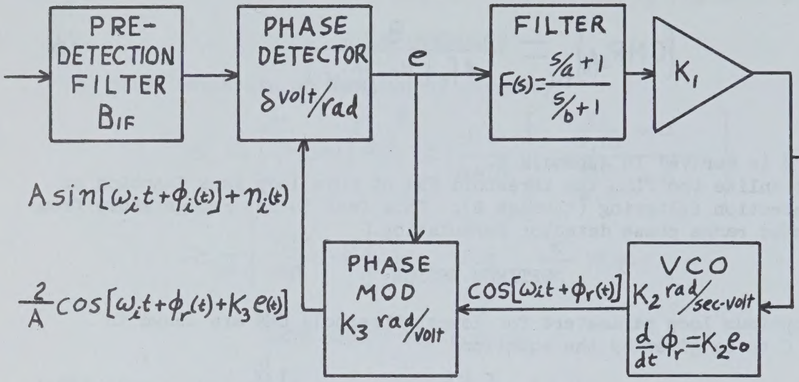

PRE-DETECTION FILTER B_{IF}
PHASE DETECTOR δ volt/rad
e
FILTER $F(s)=\frac{s/a+1}{s/b+1}$
K_1
e_o
$A\sin[\omega_i t+\phi_i(t)]+n_i(t)$
PHASE MOD K_3 rad/volt
$\cos[\omega_i t+\phi_r(t)]$
VCO K_2 rad/sec-volt
$\frac{d}{dt}\phi_r=K_2e_o$
$\frac{2}{A}\cos[\omega_i t+\phi_r(t)+K_3e(t)]$

FIG 1. Block Diagram of PLL with Phase Feedback (ERPLD)

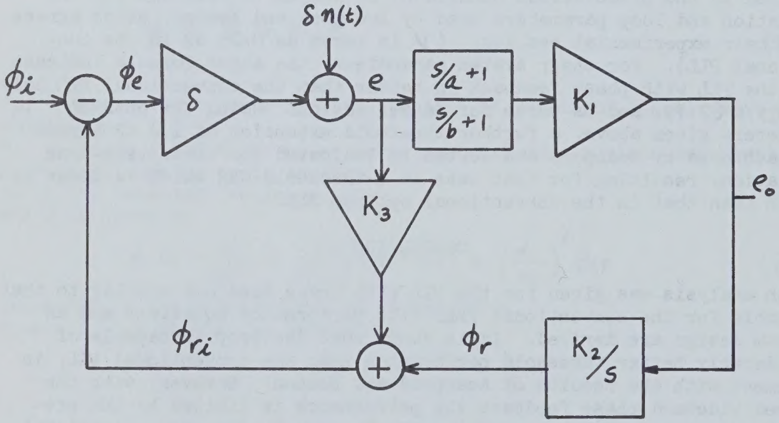

ϕ_i
$-$
ϕ_e
δ
$\delta n(t)$
$+$
e
$\frac{s/a+1}{s/b+1}$
K_1
e_o
K_3
ϕ_{ri}
$+$
ϕ_r
K_2/s

FIG 2. Linear Equivalent Diagram

Appendix A: Mean Square Phase Error

The output of the phase detector (see Fig 1), after neglecting the $2\,\omega_i$ terms, is:

$$e \doteq \delta \sin\phi_e(t) + \delta\, n(t)$$

where $\phi_e = \phi_i(t) - \phi_r(t) - K_3\, e(t)$ Phase Error

$$n(t) = -\frac{n_1(t)}{A} \sin(\phi_r + K_3 e) + \frac{n_2(t)}{A} \cos(\phi_r + K_3 e)$$

$n_1(t)$ and $n_2(t)$ are the low frequency modulations of the narrowband input noise $n_i(t)$:

$$n_i(t) = n_1(t) \sin(\omega_i t) + n_2(t) \cos(\omega_i t)$$

Also $\frac{d}{dt}\phi_r = K_2\, e_o = K f(t) * [\sin\phi_e(t) + n(t)]$

where f (t) is the impulsive response of the baseband filter and the star represents convolution.

With large loop gain and small enough noise (above threshold conditions), the phase error ϕ_e is sufficiently small to make the approximation $\sin\phi_e = \phi_e$ valid, resulting in a set of linear equations which justifies the model of Fig 2 and can be handled with the Laplace-Fourier transform calculus.

Solution of the above mesh equations in the s-domain gives

$$\Phi_e(s) = \frac{s}{KF(s) + \alpha_1 s}\,\Phi_i(s) - \frac{KF(s) + \alpha_2 s}{KF(s) + \alpha_1 s}\,N(s)$$

$$\Phi_r(s) = \frac{KF(s)}{KF(s) + \alpha_1 s}\,\Phi_i(s) + \frac{KF(s)}{KF(s) + \alpha_1 s}\,N(s)$$

where $N(s)$, $\Phi_e(s)$, $\Phi_i(s)$ and $\Phi_r(s)$ are the Laplace transforms of $n(t)$, $\phi_e(t)$, $\phi_i(t)$ and $\phi_r(t)$ respectively.

The second equation permitted us to define the closed-loop response (1).

The first equation can be written using the definition of closed loop response:

$$\Phi_e(s) = \left[\frac{1-H(s)}{\alpha_1}\right]\Phi_i(s) - \left[\frac{\alpha_2 + H(s)}{\alpha_1}\right]N(s)$$

with

$$\left[\frac{1-H(s)}{\alpha_1}\right]\Phi_i(s) = \Phi_s(s)$$

and

$$-\left[\frac{\alpha_2 + H(s)}{\alpha_1}\right]N(s) = \Phi_n(s)$$

We can transform back to the t-domain obtaining

$$\phi_e(t) = \phi_s(t) + \phi_n(t)$$

where ϕ_s and ϕ_n are the phase error components due to signal and due to noise, respectively. ϕ_s and ϕ_n are assumed to be statistically independent, therefore,

$$\overline{\phi_e^2(t)} = \overline{\phi_s^2(t)} + \overline{\phi_n^2(t)}$$

Let W_{ϕ_i} be the single-sided power spectral density of the input phase modulation. If η is the single-sided power spectral density of the noise at the input to the loop then n(t) has a power spectral density of $2\eta/A^2$ and is of bandwith $B_{IF}/2$.

We then have after using the Parseval's Theorem:

$$\overline{\phi_e^2} = \int_0^\infty \left|\frac{1-H(j\omega)}{\alpha_1}\right|^2 W_{\phi_i}\, df + \int_0^\infty \left|\frac{\alpha_2 + H(j\omega)}{\alpha_1}\right|^2 \frac{2\eta}{A^2}\, df$$

we can expand

$$|H(j\omega) + \alpha_2|^2 = |H(j\omega)|^2 + \alpha_2^2 + 2\alpha_2 \,\mathrm{Re}[H(j\omega)]$$

and get

$$\overline{\phi_e^2} = \frac{1}{\alpha_1^2}\int_0^\infty |1 - H(i\omega)|^2 W_{\phi_i}\, df + \frac{2\eta B_m}{\alpha_1^2 A^2} + \left(\frac{\alpha_2}{\alpha_1}\right)^2 \frac{\eta B_{IF}}{A^2} + \frac{4\alpha_2\eta}{\alpha_1^2 A^2}\int_0^\infty \mathrm{Re}[H(j\omega)]\, df$$

Now,

$$\int_0^\infty |1-H(j\omega)|^2 W_{\phi_i}\, df = \int_0^\infty \left|\frac{-\left(\frac{\omega}{\omega_m}\right)^2 + jb\left(\frac{\omega}{\omega_m^2}\right)}{1-\left(\frac{\omega}{\omega_m}\right)^2 + 2\xi j\left(\frac{\omega}{\omega_m}\right)}\right|^2 W_{\phi_i}\, df$$

$$\cong \int_0^{f_b} \left(\frac{\omega}{\omega_m}\right)^4 W_{\phi_i}\, df \quad , \quad if \quad 2\pi f_b = \omega_b < \omega_m ,$$

$\omega_1 > b$ (ω_1 = lowest baseband frequency in rads/sec), and $\xi < 1$. These conditions are generally met with in practice. Letting

$$Y = \int_0^{f_b} \omega^4 W_{\phi_i}\, df$$

a function of modulation only, we obtain

$$\int_0^\infty \left|1 - H(j\omega)\right|^2 df \cong \left(\frac{\alpha_1}{Kb}\right)^2 Y$$

For test-tone modulation,

$$Y_{TT} = \frac{(\Delta\omega)^2 \omega_{TT}^2}{2}$$

where $\Delta\omega$ is the peak frequency deviation, and ω_{TT} is the test-tone fequency, both in rads/sec.

For voice modulation,

$$Y_V = \omega_1 \omega_b \left(\Delta\omega_{RMS}\right)^2$$

(Y is discussed in more detail in Ref 4, Ch 6)

Similarly,

$$\int_0^\infty \mathrm{Re}\left[H(j\omega)\right] df = \int_0^\infty \frac{\omega^2\left(\frac{2\xi\omega_n}{a} - 1\right) + \omega_n^2}{\left|1 - \left(\frac{\omega}{\omega_n}\right)^2 + 2j\xi\left(\frac{\omega}{\omega_n}\right)\right|^2}\, df = \frac{Kb}{4\alpha_1 a}$$

This simple result is obtained using tabulated solutions for this type of integral (see e.g. Ref 4, Ch 2), and conventional algebraic manipulations.

Adding the above components and simplifying, we get for the mean-square value of the phase error

$$\overline{\phi_e^2} = \frac{Y}{(Kb)^2} + \frac{\eta a}{2\alpha_1^2 A^2} + \frac{\eta Kb}{2a\alpha_1^3 A^2}(\alpha_1 + \alpha_2) + \left(\frac{\alpha_2}{\alpha_1}\right)^2 \frac{\eta}{A^2} B_{IF}$$

or in more compact form

$$\overline{\phi_e^2} = \frac{Y}{(Kb)^2} + B\frac{\eta}{A^2} \tag{4}$$

where B is given by Equation 7 in the text.

Appendix B: Threshold CNR

The threshold CNR with the noise measured in a bandwith $2f_b$ is given by

$$\left[CNR_{2f_b}\right]_{TH} = \left(\frac{A^2/2}{2f_b\eta}\right)_{TH} = \frac{1}{4f_b}\left(\frac{A^2}{\eta}\right)_{TH}$$

Assuming a mean-square phase error at threshold of ν , we have from (4)

$$\left(\frac{A^2}{\eta}\right)_{TH} = \frac{B}{\nu - Y/(Kb)^2}$$

This results in

$$\left[CNR_{2f_b}\right]_{TH} = \frac{B}{4f_b\left[\nu - \frac{Y}{(Kb)^2}\right]} \tag{8}$$

Appendix C: Optimum Design:

The optimum design here is for lowest possible threshold CNR, taking into consideration both the phase error due to modulation and that due to interfering additive input noise. There are basically three parameters a, Kb and α_2. Optimization is obtained by setting the partials of (CNR_{2fb}) TH with respect to these parameters to zero.

$$\frac{\partial}{\partial a}\left[CNR_{2f_b}\right]_{TH} = 0$$

resulting in

$$a = \left[\frac{Kb\,(1+\alpha_1)}{\alpha_1}\right]^{1/2} \tag{9}$$

substitution of (9) in (8) gives

$$\left[CNR_{2f_b}\right]_{TH} = \frac{\left(\frac{\alpha_1+\alpha_2}{\alpha_1}\right)^{1/2}(Kb)^{5/2} + \alpha_2^2\, B_{IF}\,(Kb)^2}{4f_b\,\alpha_1^2\left[\nu(Kb)^2 - Y\right]}$$

This is of the form

$$\left[CNR_{2f_b}\right]_{TH} = \frac{Mx^5 + Nx^4}{\nu x^4 - Y}\,, \qquad x \equiv (Kb)^{1/2}$$

resulting in

$$\frac{\partial\left[CNR_{2f_b}\right]_{TH}}{\partial x} = 0$$

$$(Kb)^{5/2} = \frac{5Y}{\nu}(Kb)^{1/2} + \frac{4Y}{\nu}\,\alpha_2^2\, B_{IF}\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)^{1/2} \tag{10}$$

Differentiation with respect to α_2 was not performed. A digital computer solution indicates that $\alpha_2 = 1$ results in a system very close to optimum for a large range of modulation and system parameters.

References

1. A. Acampora and A. Newton, "Use of Phase Subtraction to Extend the Range of a Phase-Locked Demodulator." RCA Rev, vol 27, No. 4 PP, 577 - 599, Dec. 1966.

2. T. Watanabe and G. Sato, "Extended Range PLL Using Esaki Diode as VCO and Phase Detector." Electronics and Communications in Japan, vol. 52, no. 11, PP, 90 - 98, Nov. 1969.

3. J. Frankle, "Threshold Performance of Analog FM Demodulators." RCA Rev, vol. 27, no. 4, PP 521 - 562, Dec. 1966.

4. J. Klapper and J. Frankle, "Phase Lock and Frequency Feedback Systems: Principles and Techniques." Academic Press, New York, (scheduled for publication, Feb. 1972).

A STUDY OF A BANDWIDTH-ADAPTIVE SPEECH PROCESSOR

THOMAS E. EGER
Quest Research Corporation
McLean, Virginia

ABSTRACT

The concept of a bandwidth-adaptive speech processing system is presented. Two bandwidth measurement techniques are discussed together with the statistics gathered using each on actual speech segments. The design of the variable filter and its incorporation into the system is presented. The results of tests using actual speech segments are discussed.

I. INTRODUCTION

This paper discusses a bandwidth-adaptive speech processor which automatically adjusts the system passband to accommodate the bandwidth of input information. This processor takes advantage of the fact that speech, like most information sources, exhibits a bandwidth which varies with time. Specifically, most speech sounds are generated by one of three excitation functions: the quasi-periodic pulses which provide energy for voice sounds such as the vowels, the continuous noise excitation which result in unvoiced sounds, such as the fricatives, and finally, the single pulse excitation which provides for the unvoiced plosives. The spectra of the excitation functions themselves cover a range that includes most of the audible frequencies. However, these excitations are filtered by the vocal tract, the frequency response of which depends upon the position of the tongue, the lips, and the other articulatory organs. This response is characterized by a number of resonances or formants which shape the spectrum of the excitation function and give rise to the quality of speech sound. These response characteristics lead to spectra which are considerably different for the three types of sound excitation functions mentioned above. Generally speaking, the voiced sounds require less bandwidth than do the fricatives. Some vowels, for example, can be passed with one-fourth the bandwidth required by the fricatives. In addition, voiced sounds in English outnumber the unvoiced sounds by a factor of five to one. Taking these two facts into account, it is clear that short term spectral estimates of ordinary speech and their corresponding bandwidth will exhibit significant fluctuations leading to the conclusion that the average bandwidth of speech is considerably less than that required to pass the most wideband sounds.

In addition to bandwidth variations which occur during speech, there are also silent intervals which require no bandwidth at all. Typically, in a two person conversation, one speaker will be silent approximately half the time. In addition, there are pauses of various duration in ordinary speech. The adaptive bandwidth system is designed to take advantage of both the bandwidth variations and the silent intervals in speech.

II. ADAPTIVE BANDWIDTH TECHNIQUE

The adaptive processor is illustrated in the simplified diagram of Figure 1.0. The system represented is of the parallel-path-feed-forward type. The input information is fed to a bandwidth measurement device and through a delay to a variable filter. The estimate of the data bandwidth developed by the bandwidth measurement device is used to control the bandwidth of the variable filter through which the input data is passed. The data at the output of the variable filter thus exhibits a bandwidth which is a function of time and, under ideal conditions, is matched to the bandwidth of the input information. The output of the filter is then passed to a signal presence and gating unit. The purpose of this device is to sense the silent intervals in the speech input (or in a more general case the intervals of time during which no information is contained in the input) and gate the output to zero. The net result of the processing performed by this system is that the output is adapted to the input information both in the frequency and time domains. The variable filter adapts the channel to the bandwidth of the information in the frequency domain and the signal presence and gating unit adapts the channel to the information flow in the time domain.

There are two major problem areas which were encountered in the design of the adaptive bandwidth system. First, an acceptable bandwidth measurement device was required, and second a variable filter was needed which was capable of changing bandwidth without noticable degradation of speech quality.

III. BANDWIDTH MEASUREMENT

Two bandwidth measurement techniques were investigated, both of which provided a satisfactory estimate of the speech bandwidth. The first technique is based on a percentage power criterion and the second is a mean-squared spectrum technique. The percentage power technique estimates the bandwidth as the frequency below which a set percentage of the total waveform power is located. The Fourier transform of the input data is taken digitally using an FFT routine and the total power computed. Then the incremental power in each frequency cell is summed starting from zero until a set percentage of the total power is reached. The frequency at which this percentage is realized is used as the bandwidth measurement.

The second technique, a mean-squared spectrum estimate of bandwidth is based on the following measure,

$$\beta = \left[\frac{1}{E} \int_{-\infty}^{\infty} \omega^2 \left| F(j\omega) \right|^2 d\omega \right]^{1/2}$$

where E is the total energy in the input spectrum and $F(j\omega)$ is the Fourier transform of the input waveform. Although the implementation of this estimate was carried out on a computer, it is important to note that this bandwidth measurement has a time domain representation.

Thus,

$$\beta = \left[\frac{1}{E} \int_{-\infty}^{\infty} \left(\frac{d}{dt} f(t) \right)^2 dt \right]^{1/2}$$

where $f(t)$ and $F(j\omega)$ are a Fourier transform pair.

A fifteen second speech segment was analyzed using a bandwidth measurement unit based on the mean-squared spectrum measure. The statistics derived from this test follow a log-normal density function. Specifically,

$$p(y) = \frac{1}{\sqrt{2\pi}\,\sigma y} \exp\left[-\frac{1}{2} \left(\ln(y/m)/\sigma \right)^2 \right]$$

with

$m = 400$ Hz
$\sigma = 13$ Hz

The data are plotted in Figure 2 on a log probability plot together with a line indicating the log-normal distribution function with the aforementioned constants. Both the mean-squared spectrum estimate and the percentage power estimate resulted in satisfactory measures of speech bandwidth. In addition both techniques gave remarkably similar results.

IV. VARIABLE FILTER

The second key problem area associated with the adaptive bandwidth processor is the variable filter. The configuration illustrated in Figure 3 was selected. The input data is passed through each of the 3 filters in the filter bank and its output is switched from filter to filter depending upon the output bandwidth required. The filter bank was designed such that an input waveform having a spectrum that is approximately zero above the cutoff frequency of the filter with the smallest bandwidth, will yield identical results at the output of any of the filters. To achieve this goal it is necessary to match the phase slopes and attenuation characteristics in overlapping passbands of the various filters. With such a design, it is possible to switch from any one filter to any other filter with no phase or amplitude discontinuity, providing that both filters had sufficient bandwidth to pass the input signals. The objective of such a design concept is to eliminate any switching noise which might result from switching from filter to filter at the output of the filter bank.

Based on their "flat" amplitude response and "linear" phase response Butterworth filters were selected. Filter bandwidths of 0.5 KHz, 1.5 KHz, and 4 KHz were used. Since computer implementation was used, digital filters were designed with these bandwidths. Next phase compensation networks were designed to match the phase responses of the three filters.

The objective of the phase compensator design is a network with a constant amplitude response and a phase response which is linear. Such a network has the system transfer function represented by the delay operator:

$$A(s) = \epsilon^{-st}$$

The function A(s) has a pole at $s = -\infty$ and a zero at $s = +\infty$. The concept of a zero in the right half of the s-plane balanced by a pole symmetrically placed in the left half of the s-plane is extended to the configuration of Figure 4. This phase shifter configuration is best viewed from a geometric point of view. It is well known that the amplitude response of a network can be determined geometrically by drawing vectors from a point on the s = j axis to poles and zeros, and dividing the product of the vector magnitude to the poles. Specifically, for the configuration of Figure 4 this can be seen to be equal to:

$$\left|H(s)\right| = \frac{\left|\overline{V}_3\right| \cdot \left|\overline{V}_4\right|}{\left|\overline{V}_1\right| \cdot \left|\overline{V}_2\right|}$$

From the geometry of Figure 4 it can be seen that:

$$\left|\overline{V}_1\right| = \left|\overline{V}_3\right|$$

and $$\left|\overline{V}_2\right| = \left|\overline{V}_4\right|$$

Thus, $\left|H(s)\right| = 1$

An equally simple relationship can be derived for the phase characteristic of the compensation network. Specifically,

$$\emptyset(s) = \theta_3 + \theta_4 - \theta_1 - \theta_2$$

From the geometry involved it can be shown that

$$\emptyset(s) = 2\pi - 2\tan^{-1}\left(\frac{s-d_2}{d_1}\right) - 2\tan^{-1}\left(\frac{s+d_2}{d_1}\right)$$

It is desired that this function be linear over as great a range of s = j as possible. We begin by taking the derivative of θ. Thus,

$$\frac{d\emptyset(s)}{ds} = -\frac{2}{d_1}\left[\frac{1}{1+\frac{1}{d_1^{\,2}}\,(s-d_2)^2} + \frac{1}{1+\frac{1}{d_1^{\,2}}\,(s+d_2)^2}\right]$$

By setting the slope at $s = \pm d_2$ equal to that at zero it is possible to assure an approximately linear characteristic in the range $-d_2 \le s \le d_2$.

This implies:

$$\frac{2}{1+\left(\frac{d_2}{d_1}\right)^2} = 1 + \frac{1}{1+4\left(\frac{d_2}{d_1}\right)^2}$$

which reduces to

$$d_1 = \sqrt{2}\, d_2$$

This configuration results in a slope with maximum departure from linearity in the range $-d_2 \leq s \leq d_2$ of 2.45%, which is considered acceptable for phase compensating speech channels. Figure 5 illustrates the slopes available for various values of d_2. Only the linear range of phase response is plotted in this curve. This compensation network design was used to align the phase characteristics of the three channels. Switching from channel to channel resulted in no undesired transients so the compensation is considered satisfactory.

V. RESULTS

A computer program was written to process digitized speech segments. Software was written to estimate the input bandwidth, filter the data, and gate the output to zero during silent intervals.

An overall reduction of 32% was achieved in the bandwidth required to pass the active speech tested. If it is assumed that speech is active 50% of the time, the adaptive processor is capable of reducing the required bandwidth by at least 66%. These estimates are made for processed speech which is undistorted. Results to date indicate that if some degradation in speech quality is accepted, considerably greater reductions in bandwidth are possible.

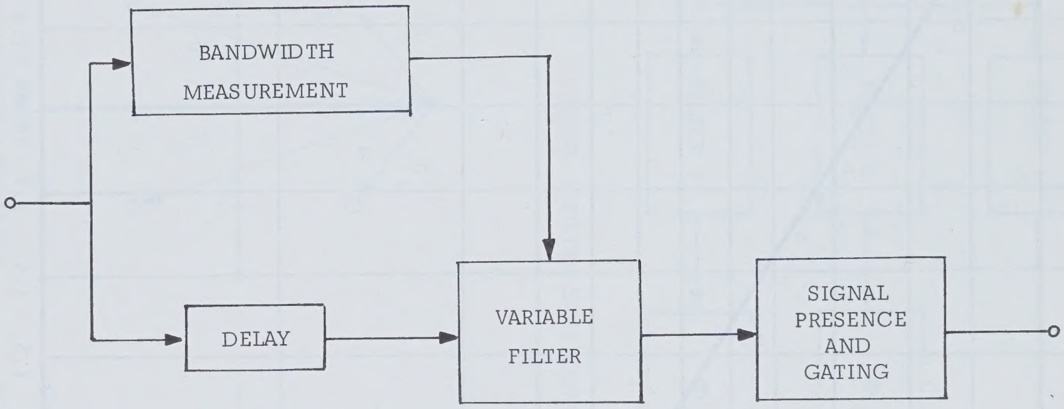

BANDWIDTH MEASUREMENT
DELAY
VARIABLE FILTER
SIGNAL PRESENCE AND GATING

FIGURE 1 ADAPTIVE BANDWIDTH SPEECH PROCESSOR

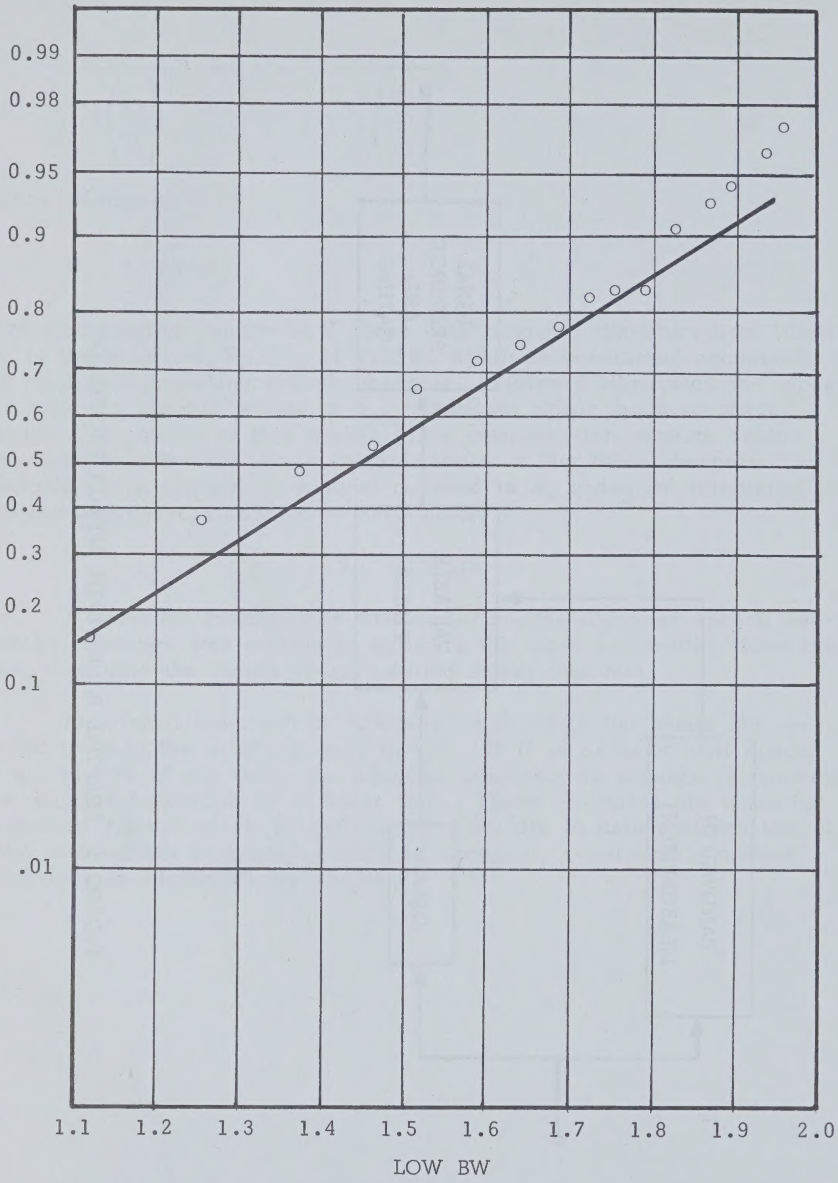


FIGURE 2 BANDWIDTH DATA (LOG PROBABILITY PLOT)

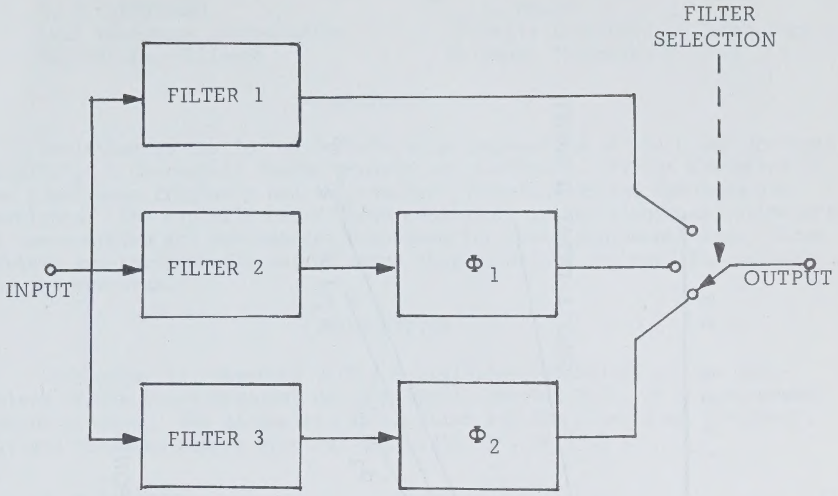

FILTER SELECTION
FILTER 1
FILTER 2
FILTER 3
INPUT
OUTPUT
Φ1
Φ2

FIGURE 3 FILTER BANK CONFIGURATION

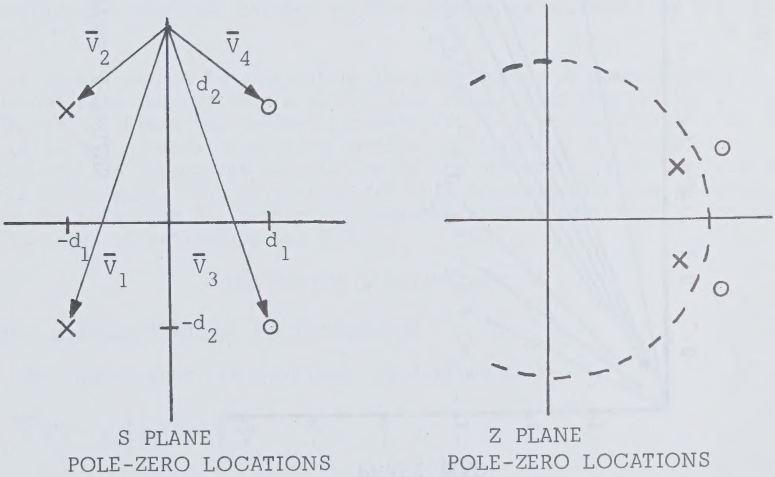

V̄2
V̄4
d2
-d1
d1
V̄1
V̄3
-d2
S PLANE
POLE-ZERO LOCATIONS
Z PLANE
POLE-ZERO LOCATIONS

FIGURE 4 PHASE COMPENSATION SYSTEM POLE LOCATIONS

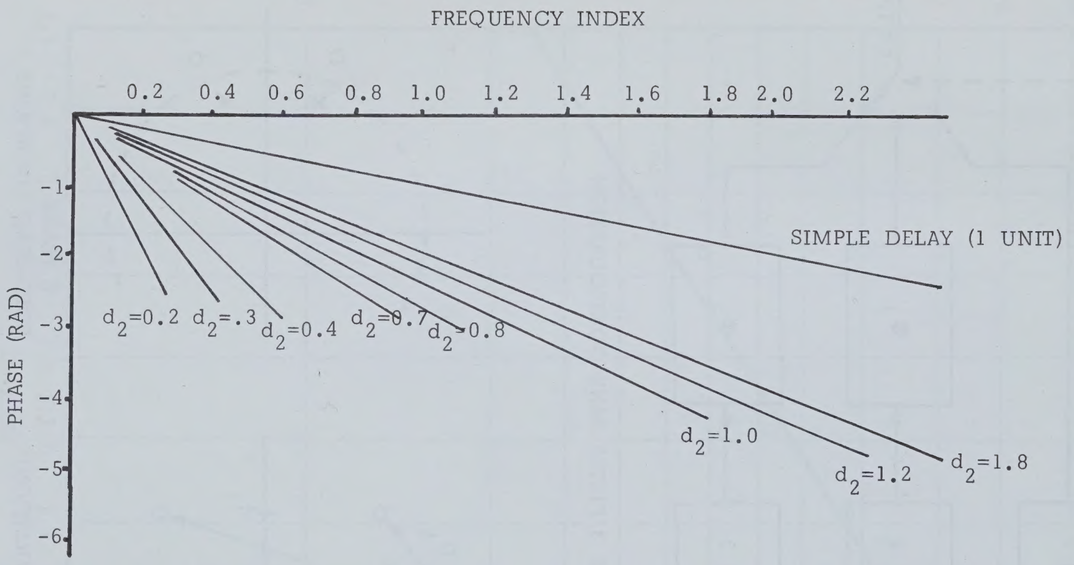


FIGURE 5 PHASE SHIFTER RESPONSE

ESTIMATION OF SPECTRAL MOMENTS

J. N. DENENBERG
Bell Telephone Laboratories
Naperville, Illinois

L. C. PEACH
Illinois Institute of Technology
Chicago, Illinois

ABSTRACT

Real-time estimation of certain shape parameters of the power spectral density of a narrowband random process is considered. Direct estimates of the power mean frequency and the standard deviation of the spectrum are developed. The equivalence of these estimators to the classical estimators is demonstrated and systems are presented for their implementation. These systems are particularly useful since they eliminate the need for estimating the spectrum.

INTRODUCTION

This paper is concerned with the real-time estimation of two parameters of the power spectral density (PSD), denoted S(f), of a narrow-band random process. The parameters in question are the power mean frequency, f_a, and the mean square spectral bandwidth, B^2, defined by

$$f_a = \frac{\int fS(f)df}{\int S(f)df} \tag{1}$$

$$B^2 = \frac{\int (f-f_a)^2 S(f)df}{\int S(f)df} \tag{2}$$

These parameters are of particular interest in radar meteorology, characterization of a fading communication channel, and radio navigation. Considerable work has been devoted to this problem as evidenced by the literature[1-14].

A direct method for estimating the parameters, or moments, of a PSD is to estimate the PSD from a set of time samples and utilize the spectral estimates to obtain the desired parameters. This is normally done by the use of A-to-D conversion of time samples, application of the Fourier Transform, and subsequent calculation of the parameter estimates from the transform outputs. The motivation for this investigation was to develop a means for obtaining the parameter estimates without resorting to the intermediate step of estimating the PSD.

DERIVATION OF ESTIMATORS

Parameter Estimation and the Periodogram

The periodogram, as used here, is defined to be

$$\hat{S}(f) = \frac{1}{T} \left| Z_T(2\pi f) \right|^2 \tag{3}$$

$$Z_T(2\pi f) = \int_t z(t)p(t)e^{-j2\pi ft}\, dt \tag{4}$$

where p(t) is a real, time limited, window function for the measurement, and z(t) is a particular sample function of the complex envelope of a narrowband random process.

Miller and Rochwarger (12) show that if f_a' and B_1^2 are the power mean frequency and mean square bandwidth, respectively, of the process after passing through the time window, then

$$f_a = f_a' - f_p \tag{5}$$

$$B^2 = B_1^2 - B_p^2 \tag{6}$$

where f_p is the mean frequency of the energy spectrum, $|P(2\pi f)|^2$, for the window function and B_p^2 is the mean square bandwidth of $|P(2\pi f)|^2$. Since $p(t)$ is real, f_p is zero and the only effect of the window function on f_a is to change the error in the estimate. B_p^2 is not zero and introduces a known bias in B^2, however, this bias can be eliminated. A problem occurs because of the time limited constraint on $p(t)$ which makes B_p^2 infinite and care must be taken in the elimination of this bias.

For the sake of convenience, the window function, $p(t)$ is normalized so that if T is the finite width of $p(t)$, then

$$\int_t p^2(t)dt = T \tag{7}$$

The window function is also separated into two factors, $\rho(t)$ and $G_T(t)$. The first factor, $\rho(t)$, is a smooth absolutely integrable function on $(-\infty, \infty)$ and has a finite mean square spectral bandwidth, B_ρ^2. $G_T(t)$ takes into account the time limited characteristic. Therefore, if

$$\xi(t) = z(t)\rho(t), \tag{8}$$

$$Z_T(2\pi f) = \int_t \xi(t)G_T(t)e^{-j2\pi ft}dt. \tag{9}$$

Thus, the effects of the window function can be included in all of the following work.

The Estimator Structure

Formulas for the parameter estimates can be obtained using the definition for each parameter and substituting the periodogram for the power spectrum. Again, care must be taken in the case of B^2 to remove the bias introduced by the window function.

The Mean Frequency Estimator Recall the definition for f_a. After substitution of the periodogram,

$$\hat{f}_a = T^{-1}\int_f f|Z_T(2\pi f)|^2 df \Big/ T^{-1}\int_f |Z_T(2\pi f)|^2 df. \tag{10}$$

The denominator can be recognized as $\hat{P}$, an estimator for the average power in the process; the numerator is $\widehat{f_a P}$, an estimator for the product of the average power with the power mean frequency.

A time domain expression for $\hat{P}$, is, of course,

$$\hat{P} = T^{-1}\int_{-T/2}^{T/2} |\xi(t)|^2 dt. \tag{11}$$

The quadrature form of this result can now be stated. Let α be the real part of ξ(t) and let β be the imaginary part of ξ(t) where the independent variable, t, of α and β is assumed understood for convenience. Then

$$\hat{P} = T^{-1} \int_{-T/2}^{T/2} [\alpha^2 + \beta^2]dt.^{*} \tag{12}$$

The same result appears in polar form as

$$\hat{P} = T^{-1} \int_{-T/2}^{T/2} A^2 \, dt \tag{13}$$

where A is the instantaneous envelope of the process after the process has passed through the window.

A similar derivation can be done for the numerator of Eq. (10). Substitution of the conjugate of the Fourier transform of z(t)p(t) for $Z_T^*(2\pi f)$ yields

$$\widehat{Pf}_a = T^{-1} \int_f f[\int_t \xi^*(t)G_T(t)e^{j2\pi ft}dt]Z_T(2\pi f)df. \tag{14}$$

which reduces to

$$\widehat{Pf}_a = -j(2\pi T)^{-1} \int_t \xi^*(t)G_T(t)[\dot{\xi}(t)G_T(t) + \xi(t)\dot{G}_T(t)]dt. \tag{15}$$

The quadrature form can be obtained by replacing ξ(t) by its real and imaginary parts, α and β, which results in

$$\widehat{Pf}_a = (2\pi T)^{-1} \int_{-T/2}^{T/2} [\alpha\dot{\beta} - \beta\dot{\alpha}]dt. \tag{16}$$

Therefore,

$$\hat{f}_a = (2\pi)^{-1} \int_{-T/2}^{T/2} [\alpha\dot{\beta} - \beta\dot{\alpha}]dt \Big/ \int_{-T/2}^{T/2} [\alpha^2 + \beta^2]dt. \tag{17}$$

In the case of $\rho(t) \equiv 1$, this result is identical to the power mean frequency estimator suggested by Bello (10). A block diagram for this measurement system is shown in Figure 1.

Some insight into this result is gained when the kernel of the numerator integral is rewritten as

$$\left\{\alpha^2 + \beta^2\right\} \cdot \left\{[1 + (\beta/\alpha)^2]^{-1} \cdot [(\alpha\dot{\beta} - \beta\dot{\alpha})\alpha^{-2}]\right\} \tag{18}$$

The first factor is A^2, the instantaneous power; the second factor is ω_i, the instantaneous frequency, where

$$\omega_i = \frac{d}{dt}[\theta(t)] = \frac{d}{dt}[\tan^{-1}(\beta/\alpha)] = 2\pi f_i \tag{19}$$

The estimator then has the intuitively consistent polar form

---

[*] Complete derivations are available in reference (14).

$$\widehat{Pf_a} = T^{-1} \int_{-T/2}^{T/2} A^2 f_i \, dt. \tag{20}$$

Note that A^2 and f_i are slowly varying functions of time. Hence, weighting the instantaneous frequency by the power at each instant of time and averaging over time yields the intuitively expected result of average power times f_a. A block diagram implementation of this envelope-frequency (or, polar form) is shown in Figure 2.

The Mean Square Bandwidth Estimator The derivation of B^2 is analogous to that for f_a. In this case, however, an additional term must be added to the definition to remove the bias introduced by the window function. The window function consists of two factors

$$p(t) = \rho(t) G_T(t). \tag{21}$$

Since $\rho(t)$ is a smooth, absolutely integrable function, the spreading of the spectrum, B_ρ^2, induced by $\rho(t)$ is finite. Therefore, B_ρ^2 can be subtracted after the estimator is found. But, B_T^2, which is introduced by the discontinuities, is not finite and must be subtracted early. Therefore, before defining $\hat{B}^2$, the spreading of the spectrum due to $G_T(t)$ must be determined.

The average value of B_T^2 can be found from its definition:

$$\overline{B_T^2} = T^{-1} \int_f f^2 \left| \int_t G_T(t) e^{-j2\pi ft} \, dt \right|^2 df. \tag{22}$$

which reduces to

$$\overline{B_T^2} = -(4\pi^2 T)^{-1} \int_t G_T(t) \ddot{G}_T(t) \, dt. \tag{23}$$

Note that the average bandwidth spreading is due to the discontinuities and is, therefore, proportional, on a particular measurement, to the percentage of power in the process at the instant of each discontinuity. This suggests that the amount of bias to be subtracted is

$$B_T^2 = -(4\pi^2 T\hat{P})^{-1} \int_t G_T(t) A^2 \ddot{G}_T(t) dt. \tag{24}$$

Hence, $\hat{B}^2$ can be defined as

$$\hat{B}^2 = (T\hat{P})^{-1} \int_f f^2 |Z_T(2\pi f)|^2 \, df - [\hat{f}_a]^2 - B_T^2 - B_\rho^2. \tag{25}$$

The quantities $\hat{P}$ and $\hat{f}_a$ have already been discussed. B_ρ^2 is, by definition,

$$B_\rho^2 = \int_f f^2 \left| \int_t \rho(t) e^{-j2\pi ft} \, dt \right|^2 df \Big/ \int_t \rho^2(t) \, dt \tag{26}$$

and, for any given window function, can be computed and used in the estimator.

Let

$$N = T^{-1} \int_f f^2 Z_T(2\pi f) Z_T^*(2\pi f) \, df - \hat{P} B_T^2. \tag{27}$$

Consideration must now be given to finding a time domain expression for N. This can be done by substituting the conjugate of the Fourier transform of $\xi(t)\, G_T(t)$ for $Z_T(2\pi f)$. It can be demonstrated that the imaginary part of N reduces to zero while the real part of N is given by

$$R[N] = (4\pi^2 T)^{-1} \int_{-T/2}^{T/2} [\dot{\alpha}^2 + \dot{\beta}^2]\, dt. \tag{28}$$

Hence,

$$\hat{B}^2 = (4\pi^2)^{-1} \frac{\int_{-T/2}^{T/2} [\dot{\alpha}^2 + \dot{\beta}^2]\, dt}{\int_{-T/2}^{T/2} [\alpha^2 + \beta^2]\, dt} - B_\rho^2 - (\hat{f}_a)^2. \tag{29}$$

This is proportional to Bello's estimator $\hat{D}_1^2$ for the mean square bandwidth when $\rho(t) = 1$.[10] This implementation is shown in Figure 3.

The polar form for $\hat{B}^2$ is also of interest and can be derived from the quadrature form by substitution of

$$\alpha = A \cos[\theta(t)] \tag{30}$$

and

$$\beta = A \sin[\theta(t)]. \tag{31}$$

Recalling that

$$\dot{\theta}(t) = \omega_i, \tag{32}$$

the result is

$$\hat{B}^2 = (4\pi^2)^{-1} \frac{\int_{-T/2}^{T/2} [A^2 \omega_i^2 + \dot{A}^2]\, dt}{\int_{-T/2}^{T/2} A^2\, dt} - B_\rho^2 - (\hat{f}_a)^2. \tag{33}$$

This form can be implemented as shown in Figure 4 and will give the same result as the quadrature form of the estimator.

Some insight can be obtained from the polar form of $\hat{B}^2$. The first term in the integral of the numerator can be attributed to the quasi-stationary, random frequency modulation of the process. The second term is due to the random amplitude modulation and is proportional to Bello's estimator $\hat{D}_2^2$ in the special case of $\rho(t) = 1$ and a gaussian process.

These results indicate that $\hat{f}_a$ and $\hat{B}^2$ (or, Bello's estimators $\hat{f}_a$ and $\hat{D}_1$ for a rectangular window function) are consistent with the classical estimators when the periodogram is used for S(f). Therefore, an error analysis of the classical estimators is also applicable to these direct methods. (14)

CONCLUSIONS

Estimators have been developed for the power mean frequency and the mean square spectral bandwidth which have the following attributes. They are consistent with the classical estimators which utilize an estimate of the spectrum. Because of this consistency, variances of these estimators can be obtained from an analysis of the variances of the classical estimators. Finally, these estimators do not require estimation of the spectrum and are, therefore, simple to implement.

BIBLIOGRAPHY

1. Schultheis, P. M., Wogrin, C. A., and Zweig, F., "Short-Time Frequency Measurement of Narrow-Band Random Signals in the Presence of Wide Band Noise". Journal of Applied Physics, Vol. 25, No. 8, August 1954, pp. 1025-1036.
2. Steinberg, H., Schultheis, P. M., Wogrin, C. A., and Zweig, F., "Short-Time Frequency Measurement of Narrow-Band Random Signals by Means of a Zero Counting Process". Journal of Applied Physics, Vol. 26, No. 2, February 1955, pp. 195-201.
3. Berger, F. B., "The Nature of Doppler Velocity Measurement". Inst. of Radio Engineers Trans. on Aeronautical and Navigational Electronics, Vol. 4, No. 3, September 1957, pp. 103-112.
4. Blackman, R. B. and Tukey, W. J., The Measurement of Power Spectra from the Point of View of Communications Engineering, Dover Publications, Inc., New York, 1959. Originally in: Bell System Technical Journal, Vol. 37, January 1958 and March 1958.
5. Pickard, T. B., "The Effect of Noise upon a Method of Frequency Measurement". Inst. of Radio Engineers Trans. on Information Theory, Vol. 4, No. 2, June 1958, pp. 83-88.
6. Swerling, P. B., "The Estimation of Doppler Shifts in Noise Spectra". Inst. of Radio Engineers Convention Record, Vol. 8, Pt. 4, 1960, pp. 148-152.
7. Mullen, J. A., "Optimal Filtering for Radar Doppler Navigators". Inst. of Radio Engineers Convention Record, Vol. 10, Pt. 5, 1962, pp. 40-47.
8. Levin, M.J., "Power Spectrum Parameter Estimation". Inst. of Electrical and Electronic Engineers Trans. on Information Theory, Vol. 11, No. 1, January 1965, pp. 100-107.
9. Ehrman, L., "Analysis of a Zero-Crossing Frequency Discriminator with Random Inputs". Inst. of Electrical and Electronic Engineers Trans. on Aerospace and Navigational Electronics, Vol. 12, No. 2, June 1965, pp. 113-119.
10. Bello, P. A., "Some Techniques for the Instantaneous Real-Time Measurement of Multipath and Doppler Spread". Inst. of Electrical and Electronic Engineers Trans. on Communication Technology, Vol. 13, No. 3, September 1965, pp. 185-292.
11. Bello, P. A., "Characterization of Randomly Time-Variant Linear Channels". Inst. of Electrical and Electronic Engineers Trans. on Communication Systems, Vol. CS-11, December 1963, pp. 360-393.
12. Miller, M. S. and M. M. Rochwarger, "On Estimating Spectral Moments in the Presence of Colored Noise". Inst. of Electrical and Electronic Engineers Trans. on Information Theory, Vol. IT-16, No. 3, May 1970, pp. 303-309.
13. Serafin, R. J. and L. C. Peach, "Calibration of Vertically Pointing Doppler Radar from Meteorological Echoes". Proc. of the 14th Radar Meteorological Conference. Univ. of Arizona, Tucson, Ariz., Nov. 1970, pp. 174-175.
14. Denenberg, J. N., "The Estimation of Spectral Moments", Laboratory for Atmospheric Probing, Department of Electrical Engineering, Illinois Institute of Technology, Chicago, Illinois 60616.

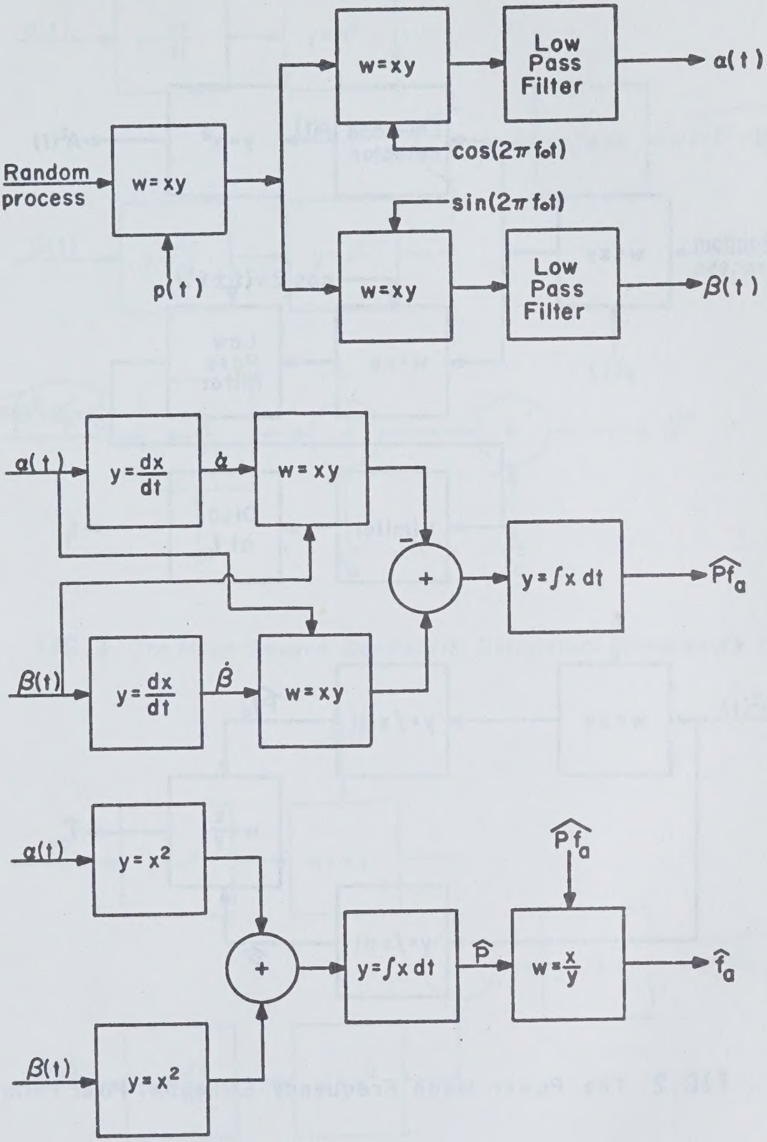

Random process
w=xy
p(t)
w=xy
Low Pass Filter
α(t)
cos(2π f₀t)
sin(2π f₀t)
w=xy
Low Pass Filter
β(t)
α(t)
y=dx/dt
α̇
w=xy
−
+
y=∫x dt
P̂f_a
β(t)
y=dx/dt
β̇
w=xy
α(t)
y=x²
+
y=∫x dt
P̂
P̂f_a
w=x/y
f̂_a
β(t)
y=x²

FIG. 1 The Power Mean Frequency Estimator: Quadrature Form

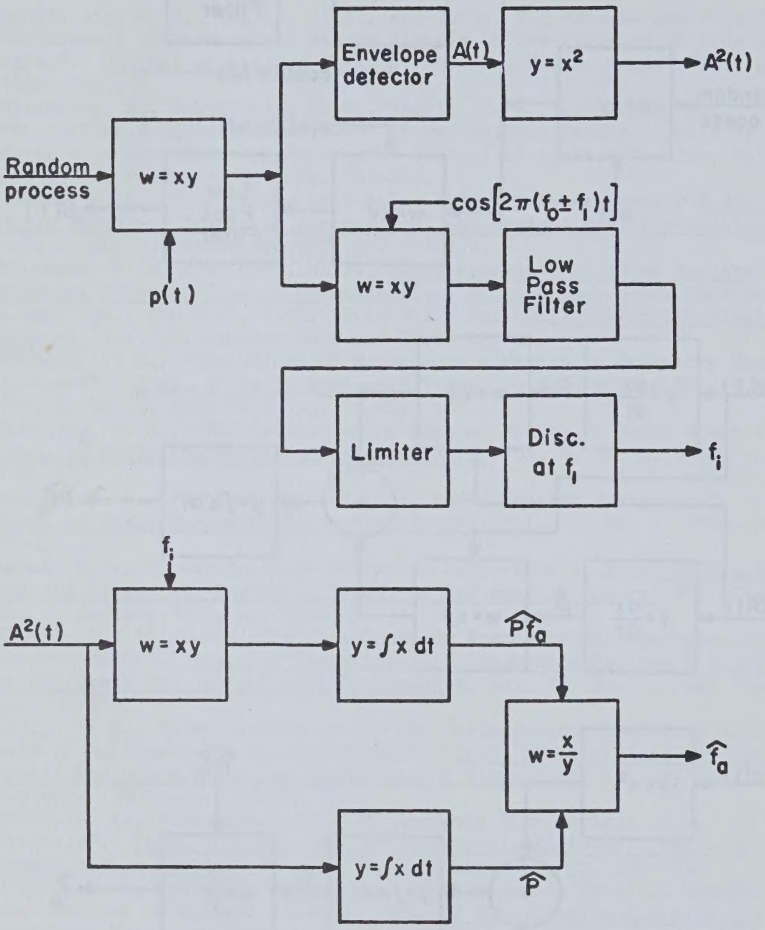

Envelope detector
A(t)
y = x²
A²(t)
Random process
w = xy
p(t)
cos[2π(f₀ ± f₁)t]
w = xy
Low Pass Filter
Limiter
Disc. at f₁
fᵢ
fᵢ
A²(t)
w = xy
y = ∫x dt
P̂fₐ
w = x/y
f̂ₐ
y = ∫x dt
P̂

FIG. 2 The Power Mean Frequency Estimator: Polar Form

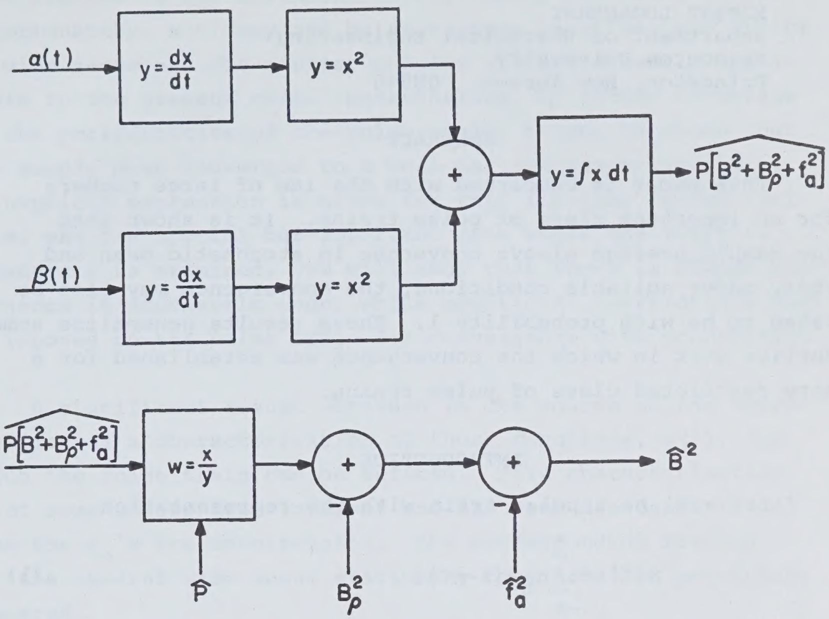

α(t)
y = dx/dt
y = x²
+
y = ∫x dt
P[B² + B_ρ² + f_a²]
β(t)
y = dx/dt
y = x²
P[B² + B_ρ² + f_a²]
w = x/y
+
−
+
−
B̂²
P̂
B_ρ²
f̂_a²

FIG.3 The Mean Square Bandwidth Estimator: Quadrature Form

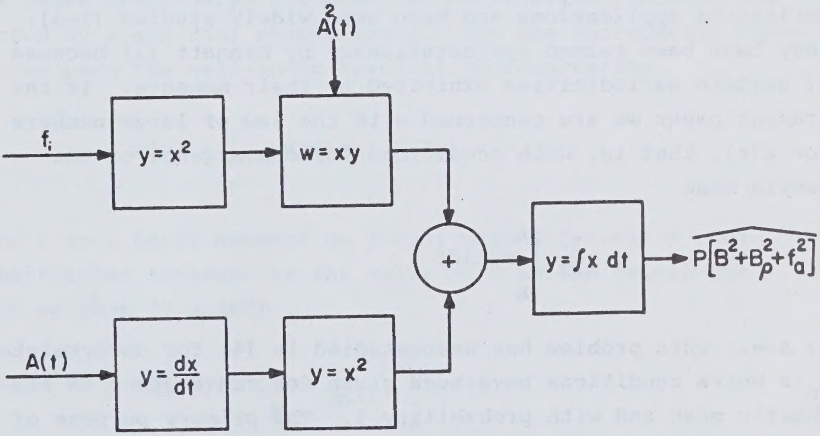

A²(t)
f_i
y = x²
w = xy
+
y = ∫x dt
P[B² + B_ρ² + f_a²]
A(t)
y = dx/dt
y = x²

FIG.4 The Mean Square Bandwidth Estimator: Polar Form

THE LAW OF LARGE NUMBERS FOR CYCLOSTATIONARY PULSE TRAINS

ROBERT LUGANNANI
Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08540

ABSTRACT

This paper is concerned with the law of large numbers for an important class of pulse trains. It is shown that the sample average always converges in stochastic mean and that, under suitable conditions, the convergence may also be taken to be with probability 1. These results generalize some earlier work in which the convergence was established for a more restricted class of pulse trains.

INTRODUCTION

Let x(t) be a pulse train with the representation

$$x(t) = \sum_{-\infty}^{\infty} \alpha_n a(t-nT). \qquad (1)$$

Here $\{\alpha_n;\ n=0, \pm 1, \ldots\}$ is a wide sense stationary process, a(t) is a nonrandom function (pulse), and T is the spacing between adjacent pulses. Pulse trains of this type arise in a variety of applications and have been widely studied [1-4]; they have been termed cyclostationary by Bennett [1] because of certain periodicities exhibited by their moments. In the present paper we are concerned with the law of large numbers for x(t), that is, with conditions for convergence of the sample mean

$$\frac{1}{2A} \int_{-A}^{A} x(t)dt$$

as $A \to \infty$. This problem has been studied in [4] for uncorrelated α_n's where conditions have been given for convergence in stochastic mean and with probability 1. The primary purpose of the following investigation is to extend these results to pulse trains for which the α_n's are merely wide sense stationary.

Most earlier studies of the law of large numbers have been limited to the treatment of stationary processes [5-7]. Unfortunately, x(t) may not be stationary in either the strict or wide sense and the earlier results are not directly applicable to the present case. Nevertheless, by taking advantage of the periodicities of the pulse train, it can be shown that the sample mean converges to a well-defined random variable. An explicit expression is given for this limiting random variable, and the special but important case where the limit is degenerate is examined. We will show that there is always convergence in stochastic mean, while additional restrictions must be imposed on the pulse train for convergence with probability 1.

A significant result obtained in the course of the investigation is a characterization of those functions, a(t), for which the pulse train can be defined. This characterization is of some independent interest and has been discussed in [4] when the α_n's are uncorrelated. The corresponding results for the general wide sense stationary case have not previously appeared.

CONVERGENCE OF x(t)

The process $\{\alpha_n; n=0, \pm 1, \ldots\}$ is assumed to be real and wide sense stationary; its mean and correlation function are denoted by μ and r(n) respectively. For the correlation function we have the well-known spectral representation

$$r(n) = \int_{-\pi}^{\pi} \exp[in\omega]\, S(d\omega)$$

where S is a Borel measure on $[-\pi,\pi)$ with $S([-\pi,\pi)) = r(0)$. Of particular interest is the value of S at the origin for which we have [5,p.489]

$$S(\{0\}) = \lim_{N} \frac{1}{2N+1} \sum_{-N}^{N} r(n).$$

Let B be an arbitrary Borel set of $[-\pi,\pi)$ and define

$$S_1(B) = \begin{cases} S(B) - S(\{0\}), & \{0\} \in B \\ S(B), & \{0\} \notin B \end{cases}$$

and

$$S_2(B) = \begin{cases} S(\{0\}), & \{0\} \in B \\ 0\ , & \{0\} \notin B \end{cases}$$

These two measures will play an important role in the following. Clearly $S = S_1 + S_2$ and $\{0\}$ is not an atom of S_1. It will be convenient to regard both S_1 and S_2 as defined and periodic (with period 2π) on $(-\infty,\infty)$.

Consider the random variable $\tilde{\alpha}$ defined by

$$\tilde{\alpha} = \underset{N}{\text{l.i.m.}} \ \frac{1}{2N+1} \sum_{-N}^{N} \alpha_n . \tag{2}$$

This limit is known to exist and to satisfy $E|\tilde{\alpha}|^2 = S(\{0\})$ [5,p.489]. The process $\{\alpha_n - \tilde{\alpha};\ n=0, \pm 1,\ldots\}$ is also wide sense stationary and, for every n, $(\alpha_n - \tilde{\alpha})$ and $\tilde{\alpha}$ are orthogonal. The spectral measure corresponding to $(\alpha_n - \tilde{\alpha})$ is easily seen to be S_1.

A special but important class of processes are those that are composed of uncorrelated random variables. Denoting the variance of the process by σ^2 we will have

$$r(n) = \begin{cases} \mu^2 + \sigma^2 & , n = 0 \\ \mu^2 & , n \neq 0. \end{cases}$$

The corresponding decomposition of the spectral measure appears as

$$S_1(B) = \frac{\sigma^2}{2\pi}\, m(B)$$

$$S_2(B) = \begin{cases} \mu^2 & , \ \{0\} \in B \\ 0 & , \ \{0\} \notin B \end{cases}$$

with m equal to Lebesgue measure. The limit $\tilde{\alpha}$ is degenerate in this case and equal to the mean.

Conditions on a(t) for convergence of the series representing x(t) have been given in [4] for a process of uncorrelated random variables. However, these conditions are no longer sufficient and a more detailed examination is required. Denote by $\mathfrak{M}$ the linear manifold of real-valued functions a(t) for which a(t), a'(t), and a"(t) are absolutely integrable on $(-\infty,\infty)$. For every $a \in \mathfrak{M}$ the series

$$\sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} a(t-nT)\, r(n-m) a(t-mT) \tag{3}$$

converges uniformly on [-T/2, T/2] to a periodic function with period T. This follows directly from the integrability of a(t) and a'(t). We define a norm for the functions of $\mathfrak{M}$ by

$$\|a\|^2 = \frac{1}{T} \int_{-T/2}^{T/2} \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} a(t-nT) r(n-m) a(t-mT) dt. \tag{4}$$

The completion of $\mathfrak{M}$ with respect to this norm will be denoted by $L_2(r)$ and two functions of $L_2(r)$ will be regarded as equal if $\|a_1 - a_2\| = 0$. With this definition of equality, $L_2(r)$ is a Hilbert space. We will examine this space in more detail presently; first, however, we state a theorem that establishes the convergence of the series (1) for all $a(t) \in L_2(r)$ and provides the justification for all sample function integrals used in the sequal. Membership in the class $L_2(r)$ represents the weakest conditions on a(t) that result in a meaningful second-order theory for x(t).

Suppose that a(t) belongs to $L_2(r)$ and that $\{a_N(t)\}$ is a sequence in $\mathfrak{M}$ that converges in norm to a(t). For every N, the series

$$x_N(t) = \sum_{-\infty}^{\infty} \alpha_n a_N(t-nT)$$

converges in stochastic mean uniformly on [-T/2,T/2]. This follows immediately from the uniform convergence of (3). We now have

THEOREM 1. The sequence $\{x_N(t)\}$ converges to a limit x(t) in the sense that

$$\lim_{N} \int_{I} E|x(t) - x_N(t)|^2 dt = 0 \tag{5}$$

for every bounded interval I. *In addition, the sample functions of* x(t) *are square integrable over every bounded interval of* $(-\infty,\infty)$ *with probability 1.*

Proof. Using the periodicity of the series (3) we have, for every integer p,

$$\int_{(p-1/2)T}^{(p+1/2)T} E|x_N(t) - x_M(t)|^2 dt = \int_{-T/2}^{T/2} E|x_N(t) - x_M(t)|^2 dt$$

$$= T \cdot \|a_N - a_M\|^2 .$$

Since $\{a_N(t)\}$ converges in $L_2(r)$, the right side of this expression goes to zero as $N,M \to \infty$. Consequently $\{x_N(t)\}$ is a Cauchy sequence in the above norm and we denote its unique limit by x(t). Since p is arbitrary, (5) clearly holds for every bounded interval. We may always select a measurable version of x(t) and it will be assumed that this has been done. Using (5) and the periodicity again we obtain

$$\int_{(p-1/2)T}^{(p+1/2)T} E|\, x(t)|^2 dt < \infty$$

and it follows (Fubini) that the sample functions of x(t) are square integrable over every bounded interval with probability 1.

We will now define a Fourier transform for the functions of $L_2(r)$. This will be useful in the ensuing discussion and will further clarify the restrictions implied by membership in $L_2(r)$. The procedure closely parallels the method used to define the transform in the case of square integrable functions.

Let $\mathfrak{H}$ denote the class of functions that are Fourier transforms of functions of $\mathfrak{M}$. Clearly $\mathfrak{H}$ is a linear manifold and, because the transforms have well-defined inverses, $\mathfrak{M}$ and $\mathfrak{H}$ are isomorphic. Now consider the Hilbert space $L_2(S)$ consisting of those functions that are square integrable with respect to the measure S. The norm in $L_2(S)$ will be given by

$$\|A\|^2 = \frac{1}{T^2} \int_{-\infty}^{\infty} |A(\omega/T)|^2 \, S(d\omega). \tag{6}$$

It can be shown that $\mathfrak{H}$ is dense in $L_2(S)$ but we omit the proof due to space limitations. If $a(t)$ belongs to $\mathfrak{M}$ and $A(\omega)$ is the corresponding function in $\mathfrak{H}$, we will have $\|a\| = \|A\|$ so that $\mathfrak{M}$ and $\mathfrak{H}$ are isometric as well as isomorphic. Since both manifolds are dense in their respective spaces, this isometric isomorphism can be extended to the whole of $L_2(r)$ and $L_2(S)$. The unique function in $L_2(S)$ corresponding to $a(t)$ in $L_2(r)$ will be called the Fourier transform of $a(t)$ and denoted by $A(\omega)$. We may thus describe $L_2(r)$ as the class of functions whose 'Fourier' transforms belong to $L_2(S)$.

If the process $\{\alpha_n, n=0, \pm 1, \ldots\}$ is one of uncorrelated random variables, the space $L_2(r)$ consists of those functions that are square integrable on $(-\infty,\infty)$ and for which the series $\sum_{-\infty}^{\infty} a(t-nT)$ converges uniformly on $[-T/2, T/2]$. These conditions have been discussed in detail in [4].

It is readily established that for almost every t

$$E\,x(t+T) = E\,x(t) \tag{7}$$

and for every τ and almost every t

$$E\,x(t+T+\tau)\,x(t+T) = E\,x(t+\tau)\,x(t). \tag{8}$$

This periodicity plays an important role in the convergence of the sample mean.

We conclude this section by introducing a pulse train $x_1(t)$ which is closely related to $x(t)$ and which will be used in the following. This new pulse train is given formally by

$$x_1(t) = \sum_{-\infty}^{\infty} (\alpha_n - \tilde{\alpha})\, a(t-nT). \tag{9}$$

The convergence of $x_1(t)$ is treated in precisely the same manner as it was for $x(t)$; the only change is that the spectral measure S_1 is used rather than S. From the inequality $S_1(B) \leq S(B)$, it follows that whenever $x(t)$ is well defined, $x_1(t)$ will also be well defined. Clearly $x_1(t)$ has the same periodicity properties as $x(t)$ and we have the relationship

$$C(\tau) = \lim_{A} \frac{1}{2A} \int_{-A}^{A} E\, x_1(t+\tau)\, x_1(t)dt \tag{10}$$

$$= \frac{1}{T} \int_{-T/2}^{T/2} E\, x_1(t+\tau)\, x_1(t)dt$$

$$= \frac{1}{T^2} \int_{-\infty}^{\infty} |A(\omega/T)|^2 \exp[i\omega\tau/T]\, S_1(d\omega).$$

CONVERGENCE OF THE SAMPLE MEAN

We begin by discussing the limit to which the sample mean converges. This limit is given by

$$\tilde{x} = \lim_{n} \tilde{\alpha} \int_{-\infty}^{\infty} a_n(t)dt \tag{11}$$

where $\{a_n(t)\}$ is a sequence in $\mathfrak{M}$ converging to $a(t)$. That the limit exists is guaranteed by the inequality

$$E|\tilde{\alpha}|^2 \left| \int_{-\infty}^{\infty} a_n(t)dt - \int_{-\infty}^{\infty} a_m(t)dt \right|^2 \leq \|a_n - a_m\|^2$$

which shows that, whenever $E|\tilde{\alpha}|^2 > 0$, the sequence $\{\int a_n(t)dt\}$ is a Cauchy sequence and thus converges as $n \to \infty$. When $E|\tilde{\alpha}|^2 = 0$ we take the right side of (11) to be zero. In the case where $\tilde{\alpha}$ is degenerate, the above expression can be written as

$$\tilde{x} = \frac{1}{T} \int_{-T/2}^{T/2} E\, x(t)dt. \tag{12}$$

We will have this degeneracy, for example, if the a_n's are uncorrelated or if $\Sigma|r(n)| < \infty$; in the latter case $\tilde{x} = 0$.

The following lemma is stated without proof.

LEMMA

$$E \sup_{NT \le A < (N+1)T} \left| \frac{1}{2A} \int_{-A}^{A} x(t)dt - \frac{1}{2NT} \int_{-NT}^{NT} x(t)dt \right|^2 \tag{13}$$

$$\le \frac{4}{N^2 T} \int_{-T/2}^{T/2} E\, x^2(t)dt.$$

As a consequence of this lemma, we can replace the sample mean

$$\frac{1}{2A} \int_{-A}^{A} x(t)dt$$

with the expression

$$\frac{1}{2NT} \int_{-NT}^{NT} x(t)dt, \quad N=1,2,\ldots$$

in the sense that their limits agree as A→∞ and N→∞, both in stochastic mean and with probability 1. For convergence in stochastic mean, the conclusion is obvious; for convergence with probability 1, it is a consequence of the fact that the right side of (13) is a term of a convergent series, the Chebyshev inequality, and the Borel-Cantelli lemma.

Our main result for convergence in stochastic mean is contained in

THEOREM 2

$$\underset{A}{\text{l.i.m.}} \frac{1}{2A} \int_{-A}^{A} x(t)dt = \tilde{x} . \tag{14}$$

Proof. It is readily established that

$$E\left|\frac{1}{2NT} \int_{-NT}^{NT} x(t)dt - \tilde{x}\right|^2 = E\left|\frac{1}{2NT} \int_{-NT}^{NT} x_1(t)dt\right|^2.$$

Combining this result with the above lemma it is clearly sufficient to show

$$\lim_{N} E\left|\frac{1}{2NT} \int_{-NT}^{NT} x_1(t)dt\right|^2 = 0. \tag{15}$$

The essential feature of the proof is the inequality

$$E\left|\frac{1}{2NT}\int_{-NT}^{NT} x_1(t)dt\right|^2 \leq \frac{3}{2N}\, C(0) + \frac{1}{T^2}\int_{-\infty}^{\infty}\left(\frac{\sin N\omega}{N\omega}\right)^2 |A(\omega/T)|^2 S_1(d\omega) \tag{16}$$

with $C(\tau)$ given by (10). The proof of (16) is omitted due to space limitations; it is similar in many respects to the proof of the corresponding inequality in [4]. Because $S_1(\{0\}) = 0$, it follows that the integral on the right side of (16) goes to zero as $N\to\infty$, thus establishing (15).

Theorem 2 shows that there is always convergence to $\tilde{x}$ in stochastic mean. The next result is concerned with convergence with probability 1; the limit must, of course, be $\tilde{x}$ in this case also.

THEOREM 3 *If either of the equivalent conditions*

$$\frac{1}{2NT}\int_{-2NT}^{2NT}(1-|\tau|/2NT)C(\tau)d\tau \leq K(\ln N)^{-\gamma} \tag{17}$$

or

$$\frac{1}{T^2}\int_{-\infty}^{\infty}\left(\frac{\sin N\omega}{N\omega}\right)^2 |A(\omega/T)|^2 S_1(d\omega) \leq K(\ln N)^{-\gamma} \tag{18}$$

hold for N sufficiently large, where K is a fixed constant and $\gamma > 2$, *then*

$$\lim_{A} \frac{1}{2A}\int_{-A}^{A} x(t)dt = \tilde{x} \tag{19}$$

with probability 1.

The proof is lengthy and is omitted as is the proof of the following

COROLLARY. *If, for* τ_o *sufficiently large or* ω_o *sufficiently small, either*

$$\int_{|\tau|>\tau_o} |\ln|\tau||^{\gamma} \left|\frac{C(\tau)}{\tau}\right| d\tau < \infty \tag{20}$$

or

$$\int_{|\omega|\le\omega_0} |\ln|\omega||^{\gamma} \, |A(\omega/T)|^2 \, S_1(d\omega) < \infty \qquad (21)$$

is satisfied for $\gamma > 2$, *then the conclusion of Theorem 3 holds*.

This corollary shows that convergence with probability 1 depends upon the behavior of $C(\tau)$ when τ is large or, equivalently, upon the behavior of $A(\omega)$ and $S_1(\cdot)$ in a small neighborhood of the origin. A variety of other conditions can be obtained which imply either (20) or (21).

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under Grant GK-1439 and Grant GK-24187.

REFERENCES

[1] W.R. Bennett, "Statistics of Regenerative Digital Transmission," *Bell Sys. Tech. Jour*., Vol. 37, 1958, pp. 1501-1542.

[2] R.W. Lucky, J. Salz, E.J. Weldon, Jr., *Principles of Data Communication*, McGraw-Hill, New York, 1968.

[3] L. Franks, *Signal Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.

[4] R. Lugannani, "Convergence of Sample Function Averages for a Class of Pulse Trains," *Proc. Fourth Princeton Conf. on Information Sci. and Sys*., 1970, pp. 362-366.

[5] J.L. Doob, *Stochastic Processes*, Wiley, New York, 1959.

[6] I.N. Verbitskaya, "On Conditions for the Applicability of the Strong Law of Large Numbers to Wide Sense Stationary Processes," *Theor. Prob. Appl*., Vol. 11, 1966, pp. 632-636.

[7] A. Blanc-Lapierre and A. Tortrat, "Sur la loi des Grands Nombres pour les Fonctions Aléatoires Stationnaires du Second Ordre," C.R. Acad. Sci. Paris, 1968, pp. 740-743.

ON NUMERICAL CALCULATIONS OF BOUNDS OF THE CAPACITY OF A TIME-DISCRETE GAUSSIAN CHANNEL WITH TIMING JITTER

KUNG YAO
School of Engineering and Applied Science
University of California
Los Angeles, California 90024

ABSTRACT

This paper considers various numerical calculations of lower and upper bounds of the channel capacity of a time-discrete channel with timing jitter and additive Gaussian noise. These bounds on the channel capacity are derived from variational methods and are given as solutions of nonlinear integral equations. For the two cases of completely correlated and completely independent jitters, extensive analytical and numerical results (in graphical forms) are obtained.

I. INTRODUCTION

In information theory, one is interested in the statistical properties of various noises and their effect on the reduction of reliable information transfer. In particular, the problem of additive Gaussian noise channel has been examined extensively in Shannon's original as well as many other studies in information theory. However, problems in synchronization, uncertainties in timing, and timing noises are of increasing interest in communication theory. In this paper, we consider a time-discrete channel, as given in Fig. 1, in which the supposedly periodic sampling times are subject to random jitters. The input signal $S(t)$, is a sample function of a real-valued, wide-sense stationary random process bandlimited to W Hz. Samples of the signal are taken at times, $\{nT + \xi_n, -\infty < n < \infty\}$, where $T = 1/2\,W$ sec. is the Nyquist sampling interval and $\{\xi_n, -\infty < n < \infty\}$ is assumed to be a strictly stationary zero-mean random sequence. Then samples, $\{Z_n = S(nT + \xi_n), -\infty < n < \infty\}$, are transmitted over a conventional additive channel. The received random sequence is given by $\{y_n = Z_n + N_n, -\infty < n < \infty\}$ where $\{N_n, -\infty < n < \infty\}$ is assumed to be independent of $S(t)$ and is a zero-mean white Gaussian random sequence with variance N. The random sequence y_n is then fed uniformly at the Nyquist rate of 2W samples per second into a low-pass filter of bandwidth W Hz. The output waveform, $Y(t)$, is then given by $Y(t) = \Sigma \operatorname{sinc}(2W(t-nT))[y_n+N_n]$.

The capacity of the above channel is the supremum of the average mutual information rate of $Y(t)$ about $S(t)$ under the constraint of a fixed input power S_o. In a conventional additive Gaussian channel, the channel capacity is well known. However, in our model, due to the presence of the random sampling jitter, the calculation of the overall channel capacity is considerably more difficult. We first obtain a lower bound on channel capacity. This bound is tight and is attained for the completely correlated jitter statistics. Then we use similar techniques to obtain an upper bound on channel capacity. However, this bound does not appear to be a tight upper bound for large signal-to-noise ratio.

Now consider two random vectors S and Y. Let the probability density $F_{S,Y}(S,Y)$ and the Gaussian probability density $G_{S,Y}(S,Y)$ have the same joint covariance matrix. Furthermore, we assume the marginal probability density $F_S(S)$ is Gaussian. Thus, $F_S(S) = G_S(S)$. Since the entropy of S and the conditional entropy of $S|Y$ satisfy $H_{F_S}(S) = H_{G_S}(S)$ and $H_{G_{S|Y}}(S|Y) \geq H_{F_{S|Y}}(S|Y)$, then

$$I_{G_{S,Y}}(S,Y) \leq I_{\substack{F_{S,Y}\\ F_S = G_S}}(S,Y) \tag{1}$$

Suppose S(t) and Y(t) are two wide-sense stationary and stationarily related random processes with the same probability structures as that of the random vectors S and Y. From Gelfand and Yaglom [1], the average mutual information rate for jointly Gaussian processes, $I_{G_{S,Y}}(S,Y)$ is given by

$$I_{G_{S,Y}}(S,Y) = W \int_{-1/2}^{1/2} \log \left[\frac{P(\lambda/T) P_Y(\lambda/T)}{P(\lambda/T) P_Y(\lambda/T) - |P_{SY}(\lambda/T)|^2} \right] d\lambda, \qquad (2)$$

where λ is a normalized, dimensionaless frequency in [-1/2, 1/2], $P(\lambda/T)$ is the input spectral density of S(t), $P_{SY}(\lambda/T)$ is the cross spectral density of S(t) and Y(t), and $P_Y(\lambda/T)$ is the output spectral density of Y(t). In [2], spectral densities $P_{SY}(\lambda/T)$ and $P_Y(\lambda/T)$ are given in terms of the jitter characteristic function $C(\lambda/T) = E\{\exp(4\pi W \lambda \xi_n)\}$ and $P(\lambda/T)$. Since the expression in (1) is valid for the S(t) and Y(t) random processes, then

$$C_L = \sup_{P(\cdot)} I_{G_{S,Y}}(S,Y) \leq \sup_{\substack{P(\cdot) \\ F_S = G_S}} I_{F_{S,Y}}(S,Y) \leq \sup_{\substack{F_S \\ P(\cdot)}} I_{F_{S,Y}}(S,Y) = C \qquad (3)$$

where the total input power is limited to S_o watts.

Thus, C_L yields a lower bound to the channel capacity C. In order to find C_L, we find the maximum of $I_{G_{S,Y}}(S,Y)$ given by (2) subject to input power S_o. The optimum $P_o(\lambda/T)$ is obtained by a variational method. From [2], the solution for $P_o(\lambda/T)$ is nonlinear and appears to be too involved to yield an explicit solution for the general case. In Sections II and III, we consider two specific useful cases of jitter statistics. For the two cases of completely correlated and completely independent jitters, we are able to obtain extensive analytical and numerical results on C_L.

Now, we use an inequality between two average mutual informations in our cascaded channels to obtain an upper bound of the jitter channel capacity. For the jittered channel given in Fig. 1, we have from [3] that

$$I(S,Y) \leq I(Z,Y)$$

Now Y is related to Z by a conventional additive Gaussian channel. Thus, for a fixed covariance matrix, we know from [4] that

$$I_{F_{Z,Y}}(Z,Y) \leq I_{G_{Z,Y}}(Z,Y)$$

Thus,

$$C = \sup_{\substack{F_S \\ P(\cdot)}} I_{F_{S,Y}}(S,Y) \leq \sup_{\substack{F_Z \\ P(.)}} I_{F_{Z,Y}}(Z,Y) \leq \sup_{P(\cdot)} I_{G_{Z,Y}}(Z,Y) = C_U \qquad (4)$$

where $P_Z(\lambda/T)$ and $P_Y(\lambda/T)$ are given in terms of the jitter characteristic function $C(\lambda/T)$ and $P(\lambda/T)$. Since

$$I_{G_{Z,Y}}(Z,Y) = W \int_{-1/2}^{1/2} \log \left[\frac{P_Y(\lambda/T)}{\phi} \right] d\lambda \qquad (5)$$

and from the convexity of the logarithmic function, we have

$$I_{G_{Z,Y}}(Z,Y) \leq W \log \int_{-1/2}^{1/2} \frac{P_Y(\lambda/T)}{\phi} d\lambda = W \log \frac{S_o + N}{N} = C_o \qquad (6)$$

Thus, (4) and (6) show that the jittered channel capacity C is less than or equal to its upper bound C_U which is less than or equal to the unjittered Shannon channel capacity C_o.

As in the calculation of C_L, in order to obtain C_U we seek to maximize $I_{G_{Z,Y}}(Z,Y)$ given by (5) subject to an input power of S_o. We note that $I_{G_{Z,Y}}(Z,Y)$ is of the same form as $I_{G_{S,Y}}(Z,Y)$ given in (2). From [2], the general solution for $P_1(\lambda/T)$ is nonlinear and appears to be as involved as the general solution for $P_o(\lambda/T)$. In the next two sections, extensive analytical and numerical results on C_L and CU are given for the two cases of completely correlated and completely independent jitters.

II. COMPLETELY CORRELATED JITTER

In the completely correlated jitter case, the random timing jitter ξ_n equals ξ for all integral n. Thus, the sampling instants are given by $nT+\xi$, $-\infty < n < \infty$. Intuitively, completely correlated jitter implies a random bias in the uniform Nyquist sampling instants. In seeking the optimum $P_o(\lambda/T)$, let $P(\lambda/T) = (1-\theta)P_o(\lambda/T)+\theta h(\lambda/T)$, where $h(\lambda/T)$ is any input spectral density with the same power S_o. Then $I_{G_{S,Y}}(S,Y)$ given by (2) is a functional of $P(\lambda/T)$ and can be denoted by $L(P)$. Then we can solve for the interior point $P_o(\lambda/T)$ in the equation $dL(P)/d\theta\big|_{\theta=0} = 0$. Furthermore, when $P_o(\lambda/T)$ is an interior point and $d^2L(P)/d\theta^2 < 0$, then $L(P_o)$ yields a local maximum for $I_{G_{S,Y}}(S,Y)$. The maximum of $I_{G_{S,Y}}(S,Y)$ is then choosen from the largest local maxima and the boundary points of P_o.

In the completely correlated case, after considerable algebraic manipulations (see [2]), $P_o(\lambda/T)$ is given by

$$P_o(\lambda/T) \geq \tilde{P}(\lambda/T) = \phi V(\lambda)/[1-|C(\lambda/T)|^2], \tag{7}$$

where ϕ = NT and

$$V(\lambda) = \tfrac{1}{2}\{|C(\lambda/T)|^2[1+4(|C(\lambda/T)|^{-2}-1)/a_o\phi]-[2-|C(\lambda/T)|^2]\} \tag{8}$$

Since $V(\lambda)$ can be negative, in order to assure that $P_o(\lambda/T)$ is non-negative, we choose

$$P_o(\lambda/T) = [\tilde{P}(\lambda/T) + |\tilde{P}(\lambda/T)|]/2 \tag{9}$$

From [2], the positive unknown parameter $a_o\phi$ is related to $P_o(\lambda/T)$ and $P_o(\lambda/T) = 0$ whenever $a_o\phi > |C(\lambda/T)|^2$. Since $|C(\lambda/T)|^2 \leq 1$, then in the numerical calculation of the non-trivial (interior point) $P_o(\lambda/T)$, we only need to consider $a_o\phi$ in the interval from 0 to 1. By direct calculations, $d^2L(P)/d\theta^2\big|_{\theta=0} < 0$. Thus, the lower bound of the channel capacity for completely correlative jitter, C_L, can be expressed as

$$C_L = 2W \int_o^{1/2} \log \left\{\frac{2[1-|C(\lambda/T)|^2]+V(\lambda)+|V(\lambda)|}{[1-|C(\lambda/T)|^2][2+V(\lambda)+|V(\lambda)|]}\right\} d\lambda \tag{10}$$

Thus far, we have only considered the calculation of C_L. We want to find a jitter statistic such that the lower bound to the channel capacity, C_L, attains the jittered channel capacity C. Furthermore, we want to find an input signal S(t) for which the average mutual information rate of Y(t) about S(t) attains C_L. From [2], we know that S(t) and Y(t) are jointly Gaussian if and only if the jitters are completely correlated. But the equalities in (1) and (3) are valid if and only if S(t) and Y(t) are jointly Gaussian. Thus, C_L attains C only for the completely correlated case. In all other cases of jitter statistics, C_L forms a strictly lower bound on C.

Now, the channel capacity $C = C_L$ of the completely correlated jitter case is obtained by numerical calculations. In order to perform the calculations, two specific jitter probability distributions are assumed. Let ξ

be a zero-mean Gaussian random variable with variance A^2, then the jitter characteristic function is given by

$$C(\lambda/T) = E\{\exp[i4\pi W\lambda\xi]\} = \exp[-8\pi^2 W^2 A^2 \lambda^2] \tag{11}$$

Let ξ be a zero-mean uniformly distributed random variable in $[-A/2,A/2]$, then

$$C(\lambda/T) = \frac{\sin 2\pi WA\lambda}{2\pi WA\lambda} \tag{12}$$

We note that in either case, as A approaches zero, $C(\lambda/T)$ approaches 1 and the effect of the random timing jitter disappears. For each fixed signal-to-noise ratio, S_o/N, each fixed A, and each fixed $C(\lambda/T)$ given by either (11) or (12) $V(\lambda)$ given by (8) is calculated for 200 test values of $a_o\phi$ taken uniformly from 0 to 1. For each test value of $a_o\phi$, $P_o(\lambda/T)$ given by (7) - (9) is calculated. Then the integral of $P_o(\lambda/T)$ over $[-1/2,1/2]$ is constrainted to be S_oT and a 20-step Simpson's rule integral is used to find the value of $a_o\phi$. If the initial test value is not within 1% of the final test value of $a_o\phi$, we repeat the search procedure for a different initial test value of $a_o\phi$. When the optimum $V(\lambda)$ (corresponding to the optimum $P_o(\lambda/T)$) is found, the jittered channel capacity, $C = C_L$, given by (10) is evaluated by a 20-step Simpson's rule integral. In Fig. 2, the completely correlated jittered channel capacity, C, with Gaussian jitter (solid lines) and uniformly jitter (dash lines) are plotted as functions of S_o/N and A. Comments and conclusions on these graphs are given in Section IV.

III. COMPLETELY INDEPENDENT JITTER

In the completely independent jitter case, each random jitter, ξ_n, $-\infty < n < \infty$, is statistically independent of others. Thus, the sampling instants, $nT + \xi_n$, $-\infty < n < \infty$, have independent random deviations from the nominal uniform Nyquist sampling instants. Intuitively, the completely independent jitter model is suitable for the characterization of most real samplers. In this case, (similar to the completely correlated case given in Section II), after considerable algebraic manipulations, $P_o(\lambda/T)$ is given by

$$P_o(\lambda/T) \geq \tilde{P}(\lambda/T) = 1/b_o[(S_o+N)/S_o-|C(\lambda/T)|^2] - (J_o+\phi)/|C(\lambda/T)|^2 \tag{13}$$

where

$$J_o = \int_{-1/2}^{1/2} [1 - |C(\lambda/T)|^2]\, P_o(\lambda/T)\, d\lambda \tag{14}$$

$$b_o = \frac{1}{J_o+\phi} \int_{-1/2}^{1/2} \frac{|C(\lambda/T)|^2 P_o(\lambda/T)}{|C(\lambda/T)|^2 P_o(\lambda/T)+J_o+\phi}\, d\lambda \tag{15}$$

In order to assure that $P_o(\lambda/T)$ is non-negative, take

$$P_o(\lambda/T) = [\tilde{P}(\lambda/T) + |\tilde{P}(\lambda/T)|]/2 \tag{16}$$

Thus, the solution of the optimum $P_o(\lambda/T)$ reduces to finding two non-negative constants, J_o and b_o, which satisfy (14)-(16). However, from (14),

$$0 \leq J_o \leq S_oT \tag{17}$$

Furthermore, we know $P_o(\lambda/T) = 0$ whenever

$$b_o > |C(\lambda/T)|^2/(J_o+\phi)[(S_o+N)/S_o-|C(\lambda/T)|^2]$$

Thus, trivial $P_o(\lambda/T)$ is avoided when

$$0 \leq b_o \leq |C(\lambda/T)|^2/(J_o+\phi)[(S_o+N)/S_o-|C(\lambda/T)|^2] \leq S_o/N(J_o+\phi) \leq S_o/N\phi. \tag{18}$$

Consequently, in the numerical computation of $P_o(\lambda/T)$ and C_L, test values of J_o and b_o can be restricted to the interval given by (17)-(18).

In order to calculate C_L, we need to calculate $d^2L(P)/d\theta^2|_{\theta=0}$. However, unlike the previous completely correlated case, the value of $d^2L(P)/d\theta^2|_{\theta=0}$ is positive or negative depending on the jitter characteristic function. However, for arbitrarily small jitter when $C(\lambda/T)$ is close to unity as well as for the two jitter characteristic functions given by (11)-(12), then $d^2L(P)/d\theta^2|_{\theta=0} < 0$. Thus the lower bound on the completely independent jittered channel capacity, C_L, is given by

$$C_L = 2W\int_0^{1/2} \log[(P_o(\lambda/T)|C(\lambda/T)|^2+J_o+\phi)/\phi]d\lambda - W\log[1+J_o/\phi] \tag{19}$$

The numerical computation techniques described in Section II can be. Now two unknown parameters, J_o and b_o, must be solved simultaneously. Then C_L given by (20) is readily evaluated. In Fig. 3, the lower bound on the completely independent jittered channel capacity, C_L, with Gaussian jitter (solid curves) and uniform jitter (dash curves) are plotted as functions of S_o/N and A.

In the evaluation of the upper bound on channel capacity, C_U, $P_1(\lambda/T)$ is given by

$$P_1(\lambda/T) = [\hat{P}_1(\lambda/T) + |\hat{P}_1(\lambda/T)|]/2 \tag{20}$$

where

$$P_1(\lambda/T) \geq \hat{P}_1(\lambda/T) = 1/[a_1+b_1(1-|C(\lambda/T)|^2] - (J_1+\phi)/|C(\lambda/T)|^2 \tag{21}$$

$$J_1 = J_1(P_1) = \int_{-1/2}^{1/2} [1-|C(\lambda/T)|^2]P_1(\lambda/T)d\lambda, \tag{22}$$

$$b_1 = \int_{-1/2}^{1/2} \{1/[|C(\lambda/T)|^2P_1(\lambda/T) + J_1 + \phi]\}d\lambda \tag{23}$$

$$S_oT = \int_{-1/2}^{1/2} P_1(\lambda/T)\, d\lambda \tag{24}$$

This time we have three unknown parameters, J_1, a_1, and b_1, and thus three integral equations to solve simultaneously. Simple calulations show the three unknown parameters can be restricted to interval given by $0 \leq J_1 \leq S_oT$, $0 \leq a_1 \leq 1/\phi$, and $0 \leq b_1 \leq 1/\phi$.

In order to obtain C_U, direct calculation shows $d^2L(P)/d\theta^2|_{\theta=0} < 0$. Thus, the upper bound of the channel capacity for completely independent jitter, C_U, is given by

$$C_U = 2W\int_0^{1/2} \log\{[P_1(\lambda/T)|C(\lambda/T)|^2 + J_1+\phi]/\phi\}d\lambda$$

The numerical computation techniques described in Section II can be used in the evaluation of C_U. In Fig. 4, C_U for Gaussian jitter is plotted as functions of S_o/N and A. Due to the excessive computation time needed to calculate C_U, no attempt was made to obtain similar results for the uniformly jittered case. Comments and conclusions for the results appearing in Figs. 2-4 are given in Section IV.

IV. CONCLUSION

Numerical calculations of the jittered channel capacity and bounds on the jittered channel capacity are given in Fig. 2-4. In all cases, as A approaches zero, the effect of the random timing jitter vanishes and the top curves in Fig. 2-4 represent the unjittered channel capacity $C_o = W \log [1 + S_o/N]$. First, consider the completely correlated jitter case. When the jitter is Gaussianly distributed the system performances given in Fig. 2 (solid curves) are seriously degraded when $A = \sigma$ is greater than about one third of the sampling time T. In other words, when the 3σ value of the jitter exceeds the sampling time, then the system performance is limited even for large signal-to-noise ratio. When the jitter is uniformly distributive, the system performances given in Fig. 2 (dash curves) are seriously degraded when A is greater than about the sampling time T even for large signal-to-noise ratio. We note that in the Gaussianly jittered case, the variance σ^2 equals A^2 while in the uniformly jittered case, the variance σ^2 equals $A^2/12$. Thus, for the same value of A, we expect C to be larger in the uniformly jitter case than in the Gaussianly jittered case. From Fig. 2 we observe these results to be indeed true. Now, consider the completely independent case. From Fig. 3-4, we observe essentially all the qualitative conclusions made above for C in the completely correlated case are valid for C_L and C_U in the completely independent case. As noted in Section III, the numerical computations of C_L and C_U in the completely independent case are considerably more involved and less accurate than that of C in the completely corrulative case. Thus, no quantitative comparisons are made among the graphs given in the completely correlated case and completely independent case. In conclusion, in the time-discrete Gaussian channel with timing jitter, the corruptive effects of the timing errors play a role similar to that of the external additive noise in limiting the channel capacity of the system.

REFERENCES

1. I. M. Gelfand and A. M. Yaglom, "Calculation of the Amount of Information About a Random Function Contained in Another Such Function," American Mathematical Society Trans., Series 2, Vol. 12, 1959, pp. 199-246.

2. K. Yao, "On the Capacity of a Gaussian Channel with Sampling Time Error," Proc. of the Fourth Hawaii International Conference on System Sciences, Jan. 1971, pp. 601-603. Also to appear in the IEEE Trans. on Information Theory.

3. R. G. Gallager, Information Theory and Reliable Communication. New York: John Wiley, 1968, p. 26.

4. A. V. Balakrishnan, "Basic Concepts of Information Theory," Lectures on Communication Theory. New York: McGraw-Hill, 1967, pp. 198-199.

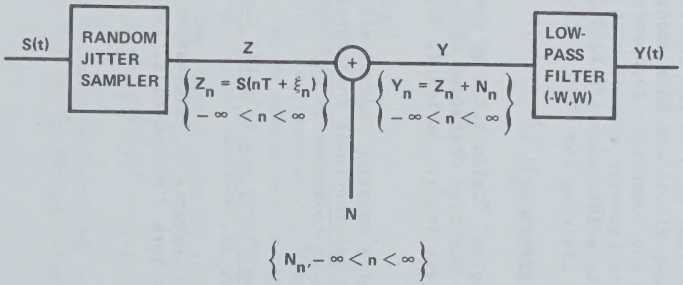

S(t)
RANDOM JITTER SAMPLER
Z
$Z_n = S(nT + \xi_n)$, $-\infty < n < \infty$
+
N
$\{N_n, -\infty < n < \infty\}$
Y
$Y_n = Z_n + N_n$, $-\infty < n < \infty$
LOW-PASS FILTER (-W,W)
Y(t)

Figure 1. Time-Discrete Channel with Random Jitter and Additive White Gaussian Noise

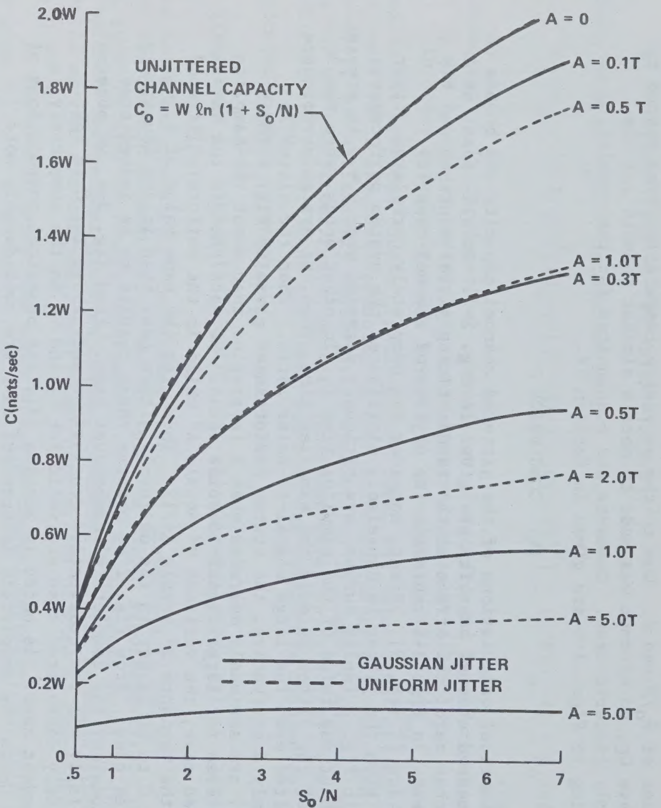


Figure 2. Completely Correlated Jitter Channel Capacity C with Gaussian and Uniform Jitters

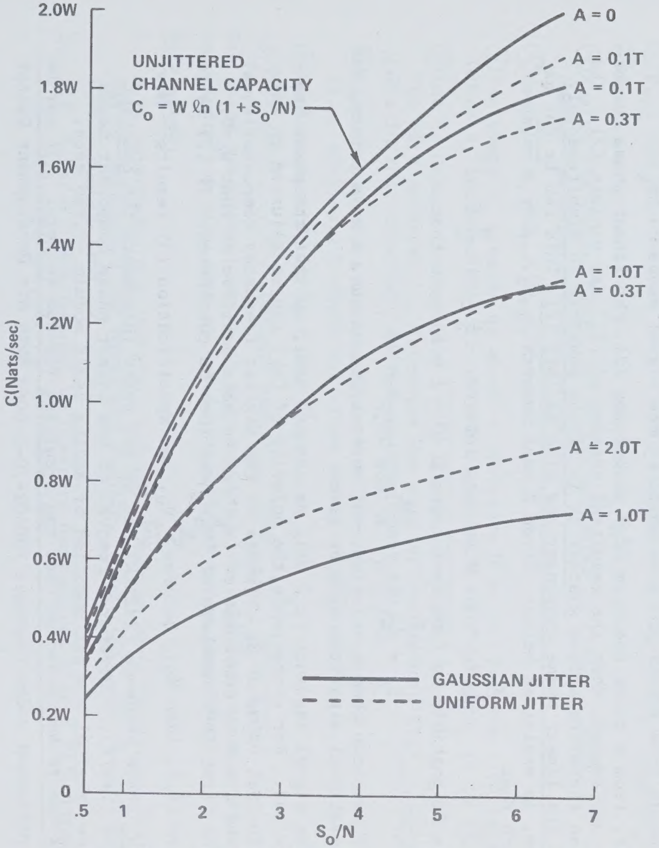


Figure 3. Lower Bound on Completely Independent Jitter Channel Capacity C_L with Gaussian and Uniform Jitters

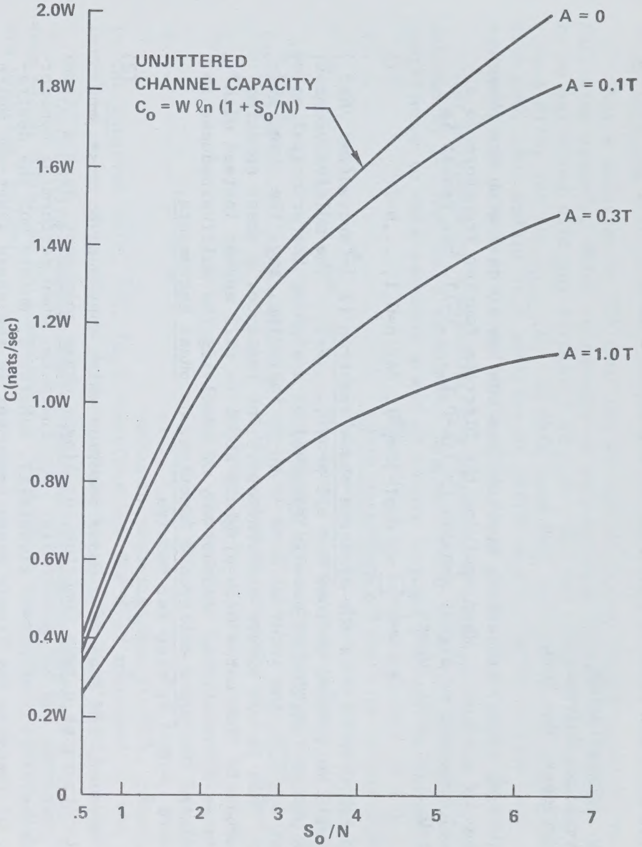


Figure 4. Upper Bound on Completely Independent Jitter Channel Capacity C_o with Gaussian Jitter

ON COMPUTING THE PHASE SPECTRUM FROM THE DISCRETE FOURIER TRANSFORM OF A SAMPLED SIGNAL*

H. SCHWARZLANDER
Syracuse University
Syracuse, New York

In some signal processing applications one has to deal with the phase spectrum of a signal. When applying the Discrete Fourier Transform to a (finite) sequence of signal samples $\{x_k\}(k=0,1,\dots,N-1)$, the result is $\{d_n\}(n=0,1,\dots,N-1)$, where

$$d_n = \frac{1}{N}\sum_{k=0}^{N-1} x_k \exp(-j2\pi\frac{nk}{N}), \text{ for } n=0,1,\dots,N-1. \qquad (1)$$

Associated with $\{d_n\}$ is the discrete phase spectrum $\{\phi_n\}=\{\arctan(\mathcal{I}m\, d_n/\mathcal{R}e\, d_n)\}$ which will be assumed defined for all $n=o,1,\dots,N-1$. The multivaluedness of the arctan function is usually resolved by assigning the principal value, so that $\phi_n \varepsilon \overline{\vdash -\pi,\pi}$, the interval from $-\pi$ to $+\pi$ including, say, the lower endpoint. This is not always satisfactory. For instance, a phase spectrum that actually fluctuates only slightly about $-\pi$ will appear instead with very large fluctuations. Another way of resolving the multivaluedness is to minimize the phase difference function, or phase increments, $\{\Psi_n\} = \{\phi_n - \phi_{n-1}\}$, that is, require

$$\Psi_n \ \varepsilon \ \overline{\vdash -\pi,\pi} \ . \qquad (2)$$

This, however, can lead to a phase spectrum which continues to rise, and thus incorporates a ramp function, or linear phase component. Such a linear phase component can be undesirable. For instance, when performing homomorphic deconvolution of echoed signals[1], Schafer[2] points out the desirability of removing any linear phase component in the phase spectrum, which corresponds to a (cyclical) time-shift of the signal sequence $\{x_k\}$.

If, from a phase spectrum $\{\phi_n\}$ satisfying (2), the linear phase component is subtracted, then the resultant version of $\{\phi_n\}$ may violate (2). The question therefore arises whether it is always possible to simultaneously remove the linear phase component and also satisfy (2). This can be put as follows, as explained below: Given K real numbers $\Psi_1,\Psi_2,\dots,\Psi_K \ \varepsilon \ \overline{\vdash -\pi,\pi}$ and such that

$$\sum_{n=1}^{K}\Psi_n = M\pi, \text{ M an \underline{even} integer, } -\frac{N}{2} \leq M < \frac{N}{2} \ ; \qquad (3)$$

then is it possible to find an integer ℓ $(0 \leq \ell < K)$ such that

$$S_\ell = \sum_{n=1}^{K}([\Psi_n + \frac{2\pi\ell}{K}]_{\bmod \overline{\vdash -\pi,\pi}}) = 0? \qquad (4)$$

It will be shown that a solution, not necessarily unique, always exists, and a computational algorithm will be given.

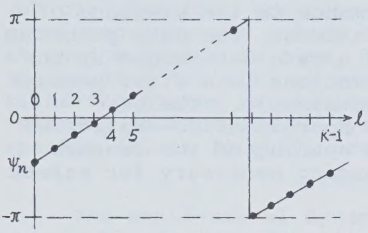
The signal sequence $\{x_k\}$ will be assumed real, so that the phase spectrum is odd. For convenience, the polarity of $\{x_k\}$ can be adjusted to make the principal value of $\phi_{N/2}$ (phase of the highest frequency component) equal zero, and the d.c. component of $\{x_k\}$ can be made positive so that $\phi_0=0$. Only the first $K=N/2$ members of $\{\phi_n\}$ need then be considered. If (2) is imposed on $\{\phi_n\}$, then $\phi_{N/2}$ becomes $\sum_{n=1}^{K}\Psi_n$, and specification (3) results. Now, a linear phase sequence $\{2\pi\ell n/K\}$ has as its phase increments the constant sequence $\{2\pi\ell/K\}$. Therefore, removal of the linear phase component means that the original Ψ's get modified by adding $2\pi\ell/K$ modulo $\overline{\vdash -\pi,\pi}$ to each, where ℓ has to be chosen so that the resulting Ψ's sum to zero. (It may be noted that in modulo $\overline{-\pi,\pi}$ arithmetic, adding $2\pi\ell/K$ to the Ψ's is equivalent to subtracting $2\pi(K-\ell)/K$.) Thence Equation 4.

---

*Work performed under Contract F30602-70-C-0060, Rome Air Development Center

That a solution to (4) need not be unique can be demonstrated by the following example, which represents an extreme case because it permits (4) to be satisfied with any integer ℓ $(0 \leq \ell < K)$: Let the Ψ's be given by $\Psi_n = \pi(2n-1-K)/K$, for $n = 1,\ldots,K$. In this case M=0, so that (4) is satisfied with $\ell=0$. But modulo $\overline{-\pi,\pi}$ addition of $2\pi\ell/K$ to all the Ψ's in this case is equivalent to a mere cyclical shift of $\{\Psi_n\}$, so that (4) is again satisfied.

For existence of a solution, the sequence of numbers $\{S_\ell\}(\ell=0,1,\ldots K-1)$ defined by (4) has to be examined. Any particular term in the summation (4) contributes to this sequence a set of values such as plotted in the Figure. Thus, $\{S_\ell\}$ is a superposition of K saw-tooth sequences of the type illustrated. This means that, from any term in $\{S_\ell\}$ to the next, there occurs either an increase of exactly 2π, a decrease by any multiple of 2π(up to $(K-1)2\pi$), or no change. (Also, $S_0=M\pi$, an even multiple of π.) Therefore, if there is no solution, that is, $S_\ell \neq 0$ for all ℓ, then either $S_\ell > 0$ for all ℓ or $S_\ell < 0$ for all ℓ; i.e., $S_\ell \geq 2\pi$ for all ℓ or $S_\ell \leq 2\pi$ for all ℓ. But each of the sawtooth sequences has its d.c. component in the interval $\overline{-\pi/K, \pi/K}$. $\{S_\ell\}$ therefore has its d.c. component in the interval $\overline{-\pi,\pi}$, so that all its values cannot lie to one side of this interval.



The following computational steps can be used to generate a phase spectrum which satisfies (4), beginning with a phase spectrum $\{\phi_n\}$ where the ϕ's are principal values.

① $\phi_{N/2}$ can only be 0 or $-\pi$. If $-\pi$, replace each ϕ_n by $(\phi_n+\pi)_{\mathrm{mod}\ \overline{-\pi,\pi}}$.

② ϕ_0 can only be 0 or $-\pi$. If $-\pi$, set $\phi_0 = 0$.

③ Assign the values $0,0,\ldots,0$ to the variables J_j $(j=1,\ldots,N/2)$.
Assign the value 0 to the variable M.

④ Compute $\Psi_n = \phi_n - \phi_{n-1}$, for $n=1,\ldots,N/2$.

⑤ For every Ψ_n greater or equal π, decrease M by 2. For every Ψ_n less than π, increase M by 2. If the final value of M is zero, go to ⑨.

⑥ Let I_j denote the interval $\overline{-\pi+(j-1)4\pi/N, -\pi+j4\pi/N}$, for $j=1,\ldots,N/2$. For each interval I_j, determine how many of the numbers $(\Psi_n)_{\mathrm{mod}\ \overline{-\pi,\pi}}$ lie in that interval, and assign that count to J_j.

⑦ Iteratively solve the equation $M + 2\ell - 2\left(\sum_{j=N/2-\ell+1}^{N/2} J_j\right) = 0$ for the first (i.e. smallest) ℓ which satisfies it.

⑧ Replace each ϕ_n by $\phi_n + 2\pi\ell\frac{2n}{N}$.

⑨ For $n=0,1,\ldots,N-2$, replace each ϕ_{n+1} by $(\phi_{n+1})_{\mathrm{mod}\ \overline{\phi_n-\pi,\phi_n+\pi}}$.

The phase spectrum obtained in this way will not only have minimum phase increments but will also be devoid of a linear phase component.

References:

1. A.V. Oppenheim, R.W. Schafer, and T.G. Stockman,Jr.: "Nonlinear Filtering of Multiplied and Convolved Signals," Proc. IEEE, 56, 8 (pp.1264-1291). August 1968.

2. R.W. Schafer: Echo Removal by Discrete Generalized Linear Filtering, Ph.D. Dissertation, M.I.T., February 1968.

REGULATION PERFORMANCE OF A CAR-FOLLOWING ALGORITHM IN THE PRESENCE OF NOISE*

V. DAVID VANDELINDE
Department of Electrical Engineering
The Johns Hopkins University
Baltimore, Maryland 21218

ABSTRACT

A car-following model based on information updating at discrete positions on the guideway has been proposed as a method of controlling high density rapid transit vehicles. This paper evaluates the system regulatory performance in the presence of additive noise in the model and observations. The data gathering scheme causes information to arrive at a rate determined by the velocity; thus, the time between observations is a first passage time whose statistics are not easily obtainable. The performance evaluation is accomplished by approximation--validated by simulation. For moderate noise the spacing of the sensors can be quite sparce compared to the spacing necessary for safety.

INTRODUCTION

One promising approach for a partial solution to the problem of moving people in and between our largest cities is automated mass transportation. If such people moving systems are successful they will have to provide service and convenience approaching that of an automobile. Thus the need includes not only large trains carrying several hundred passengers and running with headways of minutes but also much smaller vehicles with headways of only a few seconds. Once committed to the desirability of short headway systems the designer must face the more complex intervehicle control problems these present. Such problems may be separated into two classes--transient and regulation. Transient control includes all of the special problems such as merging, stopping and emergencies; while regulatory control is the normal mode--maintenance of service velocity and nominal vehicle spacing. This paper considers only the regulation problem.

While there is no unanimity among investigators as to the best means of regulation [1,2], the concensus favors the vehicle following schemes [2,3]. In this mode of regulation a vehicle attempts to maintain an optimal relationship between its state and that of the vehicles ahead of it. Such schemes may have varying degrees of inherent complexity depending on the number of preceding vehicles taken into account. The limiting case being complete awareness of the state of all other vehicles in the system. As such systems are naturally feedback, communications links must exist between vehicles. As the links become more encompassing performance naturally improves. Athans [2] has proposed the complete information system and as might be expected this works quite well. Cunningham [4] has shown that good performance continues to be obtainable if the controller

---

* This work was supported by the Urban Mass Transit Administration, under project #TRD-73.

monitors only the state of the immediately preceding vehicle (and of course the state of the vehicle in which it resides). However, even this amount of information is likely to be prohibitively expensive to obtain.

Hence the question: What is the smallest amount of information that will give satisfactory long term regulation? Clearly the answer is not zero, for errors will eventually accumulate if the vehicles are allowed to run "open-loop". Thus there must be some form of the "closed loop" control. By providing a performance evaluation of a simple data gathering system this paper partially answers the question. The system considered has sensors equally spaced on the guideway which signal the trailing vehicle when they are crossed by the leading vehicle, in effect providing accurate values of the position of the leading vehicle at random points in time, determined by the velocity profile of the leading vehicle. A regulation system using this information gathering system is examined in detail and two figures of merit relating performance to system parameters, including sensor spacing (block length), are obtained.

CAR-FOLLOWING MODEL

The small-signal dynamics of the i^{th} vehicle are assumed reasonably described by linear differential equations

$$\dot{x}^{(i)} = A\,x^{(i)} + b\,u^{(i)} + n^{(i)} \;, \tag{1}$$

where

$$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \end{pmatrix} , \quad A = \begin{pmatrix} 0 & 1 \\ 0 & -a \end{pmatrix} , \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ and } n^{(i)} = \begin{pmatrix} n_1^{(i)} \\ n_2^{(i)} \end{pmatrix} .$$

The vehicle position is x_1 and x_2 is the "velocity". We put this in quotes as a noise term, n_1, has been included in the position equation to help account for model uncertainties and therefore x_2 is not $\dot{x}_1$ as would be the case if it were true velocity. However, having made this point x_2 will be called velocity in the sequel. Control is provided by the acceleration input u and n is a noise input that is "white" and Gaussian with zero mean and correlation function $W^{(i)}\ \delta(t-s)^*$. Since vehicle mass enters only as a scale factor, it has been set to unity for convenience.

As communication exists only between a vehicle and the one following it, we need only consider such a pair and in doing so it will be convenient to define and use the relative state vector $y=x^{(i+1)}-x^{(i)}$ (no superscript is used with y as there will be only one y).

The new state equation is

$$y = Ay + b(u^{(i+1)}-u^{(i)}) + n^{(i+1)}-n^{(i)} \;. \tag{2}$$

The control $u^{(i)}$ of the trailing vehicle will be chosen to achieve the minimum of a performance criterion which is a function of the regulation error and control effort. The control function of the lead vehicle, while being chosen similarly, is unknown to the trailing controller and must be judiciously estimated. In transient control situations this task is the kernel of the control problem! For regulation the problem is much simpler, as a known nominal value exists and deviations from the nominal can be expected to be moderate and unbiased. For these reasons and to maintain mathematical tractability, let the lead vehicle's control be given as

---

* $\delta(t-s)$ is the Dirac Delta function.

$$u^{(i+1)} = f + \hat{n}^{(i+1)}, \tag{3}$$

where f is the nominal force and $\hat{n}^{(i+1)}$ is another Gaussian "white" noise term to account for deviations of the control from the nominal. It is also zero mean and has correlation function $\hat{W}^{(i+1)}\delta(t-s)$. The state equation can now be written in its final form.

$$\dot{y} = Ay - b\Delta u + n , \tag{4}$$

where $\Delta u = u^{(i)} - f$ and $n = n^{(i+1)} - n^{(i)} + b\hat{n}^{(i+1)}$.

The trailing vehicle continuously monitors its own state and acquires positional information for the lead vehicle as that vehicle passes the fixed sensors in the guideway. The scalar observation process is modeled as

$$z(\tau_m) = h\, x^{(i+1)}(\tau_m) + v_m \tag{5}$$

where τ_m is the time of the observation, $h = (1,0)$ and $\{v_m\}$ is a zero mean Gaussian sequence with correlation function $v_m\ \delta_{sm}$.* The value of the noise variance will generally be small as this noise stems from the sensors.

The cost function is taken as the standard quadratic regulator form;

$$C(t,\Delta u) = \int_t^\infty [(y-s)^T M(y-s) + \Delta u^2]dt \tag{6}$$

where s is a vector of desired states and M is a positive semi-definite symmetric weighting matrix. As y is a random process, a statistic of the cost must be chosen as the performance index. The expected value of the cost is a natural choice for the statistic and it will be used here. Letting Z(t) denote the collection of data up to t, the performance index is

$$R(t,\Delta u) = E[C(t,\Delta u)\,|\,Z(t)]. \tag{7}$$

Unlike the deterministic regulator, the stochastic regulator does not reach a limiting cost as the control interval, [t,T], increases. Thus, one must use a different figure of merit, such as cost per unit time. This quantity remains well-behaved, reaches a limit and becomes independent of initial condition. For these reasons it is one of the two figures of merit computed. The second, also closely related to the regulator problem, is perhaps more intuitively appealing; using the control law determined by the minimization of (7), the maximum expected square deviation of the state from the nominal is computed.

---

*δ_{sm} is the Kronecker delta function.

AVERAGE COST

The first figure of merit is the average cost, J_A. It is computed by a straightforward but laborious application of Dynamic Programming (5). The calculation yields

$$J_A = tr(W^{(i)}Q)+\lim_{T\to\infty} \frac{1}{T}\left\{\int_0^T tr(E(P(\sigma)M))d\sigma + \sum_{\tau_m \le T} tr[E(P(\tau_m^-))Q]\right\}, \quad (8)$$

where Q is the steady state solution to the Ricatti equation for this problem and P is the conditional covariance matrix of the state, y. We must devise a scheme for calculating the value of this expression. The task is complicated by the manner by which data is received. Because the data rate is fixed to distance rather than time, the excursions of P are not periodic, even asymptotically. The problem is solved using two approximations whose validity depends on "a" being small. Fortunately, for the type of vehicles used in mass transportation appropriate values for "a" fall in the interval .1 - .2.

The first approximation involves $P(\tau_m^+)$, the conditional covariance matrix immediately following an observation. We will assume that $P(\tau_m+)$ is a constant equal to the value it would have if the observations were equally spaced with the time between observations equal to the expected time. If observations are taken periodically in time P, quickly approaches a periodic function. The results of the second approximation will show that the variance of the time between observations, $\tau_{m+1}-\tau_m$, has a magnitude less than 2% of $\tau_{m+1}-\tau_m$. This implies that most deviations of the time between observations from the nominal will fall in the ± 20% range. This taken in conjunction with the concave nature of the covariance equations when an observation is taken, results in the value of the conditional covariance matrix immediately following an observation varying by appreciably less than 20%. In hope of improving the estimate of the possible variation in the components of the conditional covariance matrix, several data sets were selected at random from representative numbers and the covariance matrix calculated assuming periodic observations for transit time equal to the nominal and 20% either side of the nominal. The deviations and percent change are exhibited in Table 1. This table also contains a column showing the deviation of the variations from anti-symmetry in percent of the nominal. These numbers are perhaps the most significant as the covariance matrix enters approximately linearly into the expression for covariance matrix just before the next observation. Thus anti-symmetric variations will tend to cancel when expectations are taken.

Com-ponent	Nominal Value	Deviation		Max % Deviation	% Deviation from Anti-symmetry
		-20%	+20%		
P11	.486	-.011	.003	2.2	1.6
P12	.367	.042	-.037	11.4	1.4
P22	5.49	-.19	.25	4.6	1.1
P11	.248	-.002	0	0.8	0.8
P12	.192	.023	-.022	0.2	0.5
P22	9.39	-.44	.52	5.5	1.0
P11	.246	-.001	.001	0.4	0.0
P12	.103	.011	-.010	10.7	1.0
P22	3.61	-.08	.10	2.8	0.6

Table 1

Sensitivity of $P(\tau_m^+)$ to Changes in Period.

The second approximation involves the time between observations. This time, $\Delta\tau$, can be characterized as the first passage time of a second order Markov process. A consideration of this problem results in the moment generating function of $\Delta\tau$ being determined as an infinite series of products of exponential functions and Weber functions. This exact form was rejected because of its complexity in favor of a simple approximation.

One can write the mean and variance of $\Delta\tau$ for a Brownian motion process as (6)

$$E[\Delta\tau]=b/\mu \quad \text{and} \quad Var[\Delta\tau] = Var(y(E(\Delta\tau))]/\mu^2, \qquad (9)$$

where b is the length to be traversed and μ is the mean velocity. As the vehicle equations are well-damped this seemed a likely approximation. The version of these equations for the vehicle model are:

$$E[\Delta\tau]=ba/f=\overline{\Delta\tau} \text{ and } Var[\Delta\tau]= a^2(w_1\overline{\Delta\tau} + w_2(a\overline{\Delta\tau}-1+e^{-a\overline{\Delta\tau}})/a^3)/f^2, \qquad (10)$$

where $w_1 = 2W_{11}^{(i)}$ and $w_2 = 2W_{22}^{(i)} + \hat{W}_{22}^{(i+1)}$. An extensive computer simulation of the process resulted in a sample mean that was always within 2% of the approximation. As might be expected a very slight bias toward longer intervals was noted. Figure 1 + 2 show a comparison of the simulation with the approximation to the variance. Again there seems to be a slight under estimation, but not severe enough to merit further study.

Now, J_T can be calculated in a straightforward manner. If t is large enough for all processes to be in statistical steady state, J_T can be expressed as

$$J_T= tr(W_iQ)+E\left[\frac{1}{\tau_{m+1}-\tau_m}\left\{\int_{\tau_m}^{\tau_{m+1}} tr(P(\sigma)M)d\sigma+tr[(P(\tau_m^-)-P(\tau_m^+)Q]\right\}\right] \qquad (11)$$

but in the interval of integration

$$\dot{P} = AP + PA^T + W \qquad (12)$$

whose solution is

$$P(t) = \Psi(t,\tau_m^+)P(\tau_m^+)\Psi(t,\tau_m^+) + \int_{\tau_m^+}^{t} \Psi(t,\sigma)W\,\Psi^T(t,\sigma)d\sigma \ . \qquad (13)$$

For the model being considered, this calculation can be made in closed form; and since $P(\tau_m^+)$ is a computable constant, $P(t)$ can be found as an explicit function of time. The integral in (11) can be explicitly calculated, with the result involving terms linear, and exponential in $\tau_{m+1}-\tau_m$. This is also true of the last term in the expression. These expressions are then divided by $\tau_{m+1}-\tau_m$ and the indicated expectation taken. This cannot be performed exactly since we have available only the first and second order moments of the transit time, $\tau_{m+1}-\tau_m$. However, these are the only two significant moments in the expectation. The expectation is computed by rearranging the series for the exponential to form a series in the mean transit time and the central moments of the transit time. This series is then truncated after the variance term to yield the approximate expectation.

A computer program following the above outline has been written and results for several data sets are shown as Figures 3 thru 4 . Each of the figures displays average cost vs. block length with a third parameter varying to form the ensemble of curves. Notice that, except in cases of severe noise or low mean velocity the cost ratio between, say, 20 and 50 ft. blocks is only about two.

MEAN SQUARE DEVIATION

As a second figure of merit we compute a maximum expected value of the square of the deviation from the desired state relationships between the leading and trailing vehicle. This calculation is made under the assumption that the optimal control law is being followed by the trailing vehicle. Since following this control law results in the unconditional expectation of the state, y, equalling the desired state, s, we are computing the unconditional covariance matrix of y,

$$\mathrm{Var}(y) = E[\mathrm{Var}(y|Z)] + \mathrm{Var}[E(y|Z)] \ .$$

The maximum value of the covariance matrix will occur when an observation occurs as the "open loop" control interval will have just ended at this point. Unlike the conditional covariance matrix the covariance matrix is a continuous function; the discontinuities in the two contributing terms being equal and opposite.

Define $\quad P(t) = E[\mathrm{Var}(y|Z)] \qquad (14)$

and $\quad S(t) = \mathrm{Var}[E(y|Z)] \qquad (15)$

then $\quad \dot{S} = \bar{A}\,S + S\,\bar{A}^T + W^{(i)} \qquad (16)$

where $\bar{A}$ is the closed-loop plant matrix and

$$P = AP + \dot{P}\,A^T + W. \qquad (17)$$

As indicated the boundary condition on S is the opposite of the condition on P; that is, when an observation occurs

$$S(\tau_m^+) = S(\tau_m^-) - P(\tau_m^+) + P(\tau_m^-) \ . \qquad (18)$$

The minimum value of S is $S(\tau_m^-)$ and we compute this assuming the nominal value for the change in P and the length of the transit interval. Since (16) is linear the contribution due to $W^{(i)}$ can be computed separately and added to the homogeneous part of S, $\hat{S}$.

Thus

$$\hat{S}(\tau_m^-) = \Phi(\overline{T},0)(\hat{S}(\tau_m^-)+P(T)-P(0))\Phi^T(T,0) \qquad (19)$$

This is a linear equation and is easily solved by matrix inversion. The inhomogeneous contribution satisfies

$$0 = \overline{A}(S-\hat{S}) + (S-\hat{S})A^T + W^{(i)} \qquad (20)$$

since $\overline{A}$ is a stability matrix. The value of $P(\tau_m^-)$ is computed using the approximations from the computation of average cost. The final result is obtained by combining the three terms.

$$Y(\tau_m^-) = \hat{S}(\tau_m^-) + S-\hat{S} + P(\tau_m^-) \ . \qquad (21)$$

These equations were programmed and the results are shown as Figures 5 to 6. In these figures w_1 is equal to $2W_{11}^{(i)}$. $W_2^{(i)}$ equal to one and w_2 is equal to the sum of $\hat{W}_2^{(i+1)}$ and $2W_{22}^{(i)}$. The good regulation achieved is evident from the figures with the maximum root of the positional square deviation under four feet.

CONCLUSION

In this report we have examined the performance of an optimal control system for regulating intervehicle states when data is gathered by fixed guideway sensors. We were able to generate two figures of merit for the performance--average cost and mean square deviation. Both seem to indicate that the spacing between the sensors can be quite large, 50 ft. or more, without seriously degrading the regulatory performance of the control system.

REFERENCES

1. Hajdu, L.P. et al, "Design and Control Considerations for Automated Ground Transportation Systems," Proceeding of the IEEE, Vol. 56 #4, April, 1968, pp. 493-513.

2. Levine, W.S. and Athans, M., "On the Optimal Error Regulation of a String of Moving Vehicles," IEEE Transactions on Automatic Control, Vol. AC-11, #3, July, 1966, pp.355-361.

3. Bender, J.G. and Fenton, R.E. "On the Flow Capacity of Automated Highways," Transportation Science, May, 1970, pp. 52-63.

4. Cunningham, Edward, "Velocity/Spacing Regulation for a String of Moving Cars," Johns Hopkins University - Applied Physics Laboratory, Internal Memo, MCS-0-125, Aug. 7, 1969.

5. VandeLinde, V. David, "Optimal Vehicle Regulation with Discrete Positional Sensing," Internal Memo, MCS-3-192, Nov. 17, 1970.

6. Darling, D.H. and Siegert A. "The First Passage Problem for a Continuous Markov Process," Annals of Math Stat., 1953, pp. 624-639.

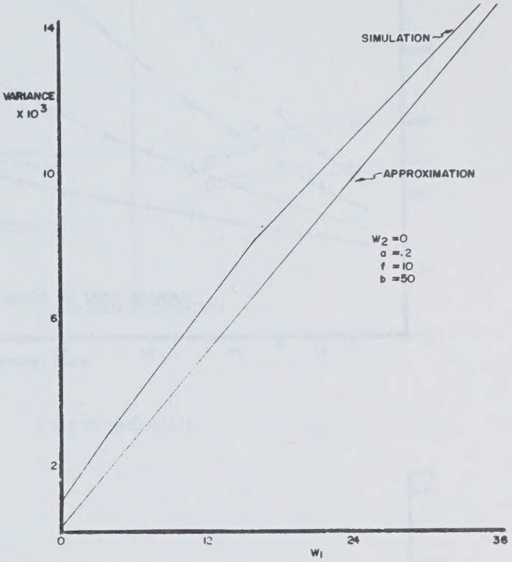


Figure 1

Simulation Compared to Approximation

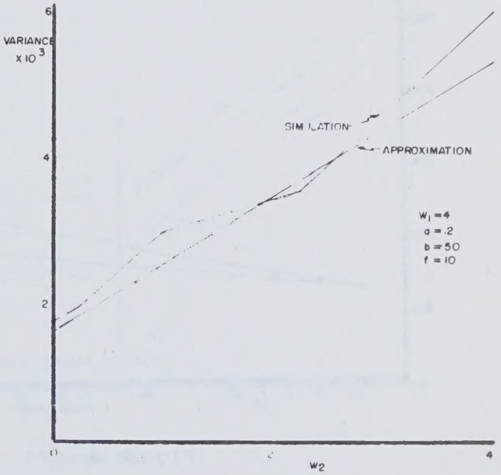


Figure 2

Simulation Compared to Approximation

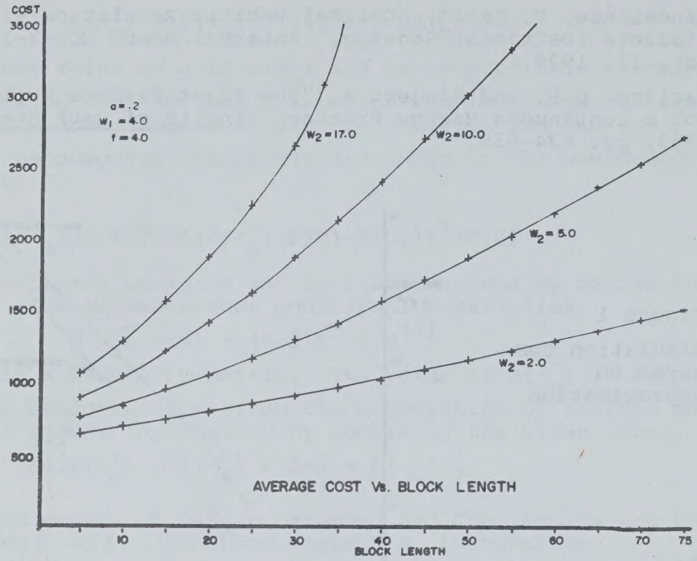


Figure 3

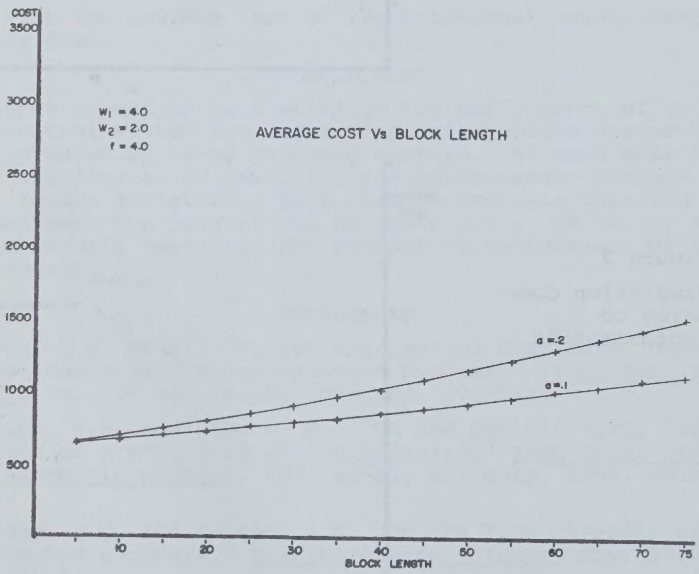


Figure 4

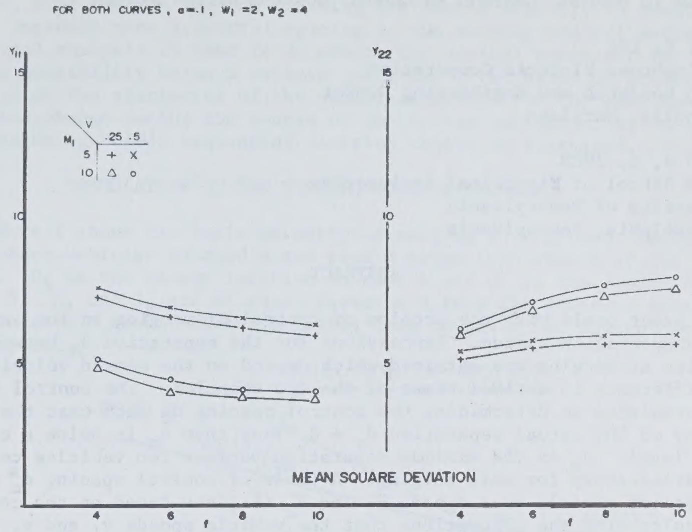


Figure 5

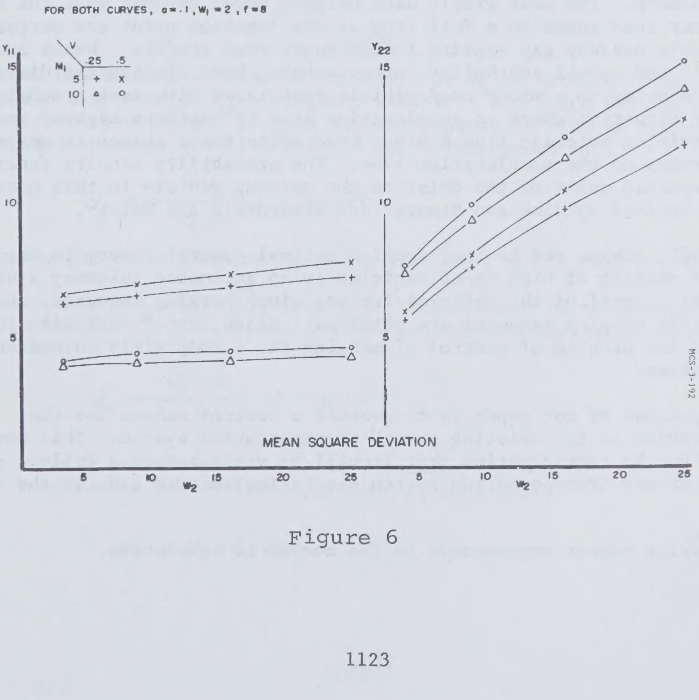


Figure 6

AN APPROACH TO MERGING CONTROL IN GROUND TRANSPORTATION SYSTEM

HENRY E. LEE
Westinghouse Electric Corporation
Ocean Research and Engineering Center
Annapolis, Maryland

DAVID W. C. SHEN
Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia, Pennsylvania

ABSTRACT

This paper deals with the problem of control of merging in the existing ground transportation system. Expressions for the separation d_s between two vehicles at merging are obtained which depend on the sensed velocities and the difference in arrival times of the two vehicles. The control problem is formulated as determining the control spacing d_c such that the probability of the actual separation $d_s + d_c$ less than d_m is below a certain specified level. d_m is the minimum separation between two vehicles considered satisfactory for safe merging. A table of control spacing d_c^* with combinations of vehicle mean speeds $\overline{v}_1$ and $\overline{v}_2$ is given based on the computational results with the assumptions that the vehicle speeds v_1 and v_2 are independently normally distributed. Considering the case where the statistics of the vehicle speeds change during the system operation, an on-line procedure using the sequential decision theory is discussed.

INTRODUCTION

The merging problem in ground transportation system has been analyzed by many authors. The most widely used merging model assumes that the driver on the minor road comes to a full stop at the junction point and merges when an acceptable headway gap appears in the major road traffic. Weiss and Maradudin[1*] and McNeil and Smith[2], among others, have studied the distribution of the delay to a minor road vehicle associated with such a model. There is a situation where an acceleration lane is beside a highway and that an arriving motorist from a minor road waits for a chance to merge as it accelerates on the acceleration lane. The probability density function and the expected value of the delay to the merging vehicle in this case have been derived by Mine and Mimura[3] and Blumenfeld and Weiss[4].

In 1967, Athans and Levine[5] applied optimal control theory to the control of merging of high speed vehicles in an automated guideway system. The optimal control of the vehicles for any given merging sequence, and the best possible merging sequence are obtained. Hajdu, etc.[6] have also investigated the problem of control of merging for a completely automated highway system.

The purpose of our paper is to present a control scheme for the merging problem in the existing ground transportation system. This work is motivated by the consideration that it will be years before a fully automated ground transportation system can be implemented and, in the

---

[*] Super-script number corresponds to the number in References.

meantime, some control for the existing traffic is urgently in need. Our present approach uses a control spacing as the merging control parameter. Statistical analysis is used to determine the control parameter for reducing the miss possibility below a certain specified level based on priori knowledge of the statistics of the vehicle speeds. For the case where these statistics change during the course of the system operation, an on-line procedure using Wald's sequential decision theory is discussed.

FORMULATION OF THE MERGING CONTROL PROBLEM

Figure 1 shows the basic geometry of merging in a ground transportation system where vehicles on road A and road B merge into road C at the junction point M. G_A is the sensor location on road A and G_B is the sensor location on road B. G_A and G_B are of equal distance d from the junction point M where distance d should be sufficiently long to allow for control actuation to take place.

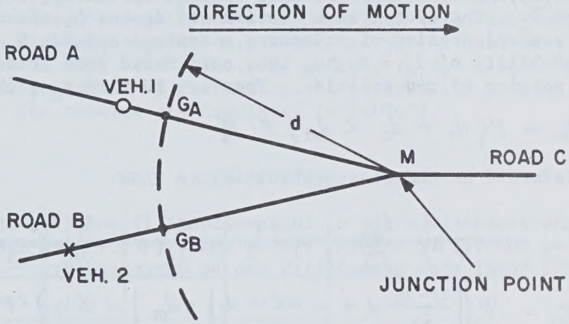


FIGURE I

Assume the time of arrival of vehicle 2 at G_B is Δt later than the time vehicle 1 arrives at G_A. Speeds of vehicle 1 and 2 are designated by v_1 and v_2 correspondingly. If $v_1 \geq v_2$, the separation between two vehicles when vehicle 2 reaches junction point M can be expressed as

$$d_s = \frac{v_1 - v_2}{v_2} \cdot d + v_1 \cdot \Delta t \tag{1}$$

Above expression for separation d_s holds also when $v_1 < v_2$ and $v_1\left(\Delta t + \frac{d}{v_2}\right) \geq d$.

When $v_1 < v_2$ and $v_1\left(\Delta t + \frac{d}{v_2}\right) < d$, vehicle 1 will reach junction point M later than vehicle 2. The separation between two vehicles when vehicle 1 reaches M is

$$d_s = \frac{v_2 - v_1}{v_1} \cdot d - v_2 \cdot \Delta t \tag{2}$$

The merging control problem is to determine the control spacing d_c which is required to add to the separation d_s such that

$$d_s + d_c \geq d_m \tag{3}$$

where d_m is the minimum separation between two vehicles considered satisfactory for safe merging. Decreasing d_m will increase the flow capacity of the system at the expense of safety of the merging process. In a fully automated system, where vehicle speeds v_1 and v_2 are deterministic, control spacing d_c can be determined from above equations (1) to (3) for each set of $\overline{v}_1$, $\overline{v}_2$ and Δt after a particular d_m is selected. Since fully automated ground transportation systems are not likely to be in operation in the near future, a statistical approach to the merging control problem seems to be more appropriate and compatible with the existing traffic.

DETERMINATION OF CONTROL SPACING FOR GIVEN VEHICLE SPEED STATISTICS

In the existing traffic, where vehicles are human-controlled, vehicle speeds v_1 and v_2 are obviously random variables. The distributions of v_1 and v_2 depend upon how closely the true vehicle speeds approximate the mean speeds $\overline{v}_1$ and $\overline{v}_2$. Therefore, in a statistical approach, one must consider the merging control problem of selecting a control spacing d_c to insure that the probability of $d_s + d_c < d_m$ does not exceed some tolerable value P_s for each merging of two vehicles. That is, to find d_c such that

$$P_m = P(d_s + d_c < d_m) \leq P_s \tag{4}$$

P_m will be referred as the miss probability.

Using the expressions for d_s in equations (1) and (2), and since cases $v_1 \geq v_2$, $v_1 < v_2$ and $v_1(\Delta t + \frac{d}{v_2}) \geq d$, $v_1 < v_2$ and $v_1(\Delta t + \frac{d}{v_2}) < d$ are mutually exclusive, the total miss probability can be expressed as

$$\begin{aligned} P_m = {} & P\left(\left|\frac{v_1 - v_2}{v_2} d + v_1 \Delta t + d_c\right| < d_m \;\middle|\; v_1 \geq v_2\right) \cdot P(v_1 \geq v_2) \\ & + P\left(\left|\frac{v_1 - v_2}{v_2} d + v_1 \Delta t + d_c\right| < d_m \;\middle|\; v_1 < v_2,\ v_1\left(\Delta t + \frac{d}{v_2}\right) \geq d\right) \\ & \cdot P\left(v_1 < v_2,\ v_1\left(\Delta t + \frac{d}{v_2}\right) \geq d\right) \\ & + P\left(\left|\frac{v_2 - v_1}{v_1} d - v_2 \Delta t - d_c\right| < d_m \;\middle|\; v_1 < v_2,\ v_1\left(\Delta t + \frac{d}{v_2}\right) < d\right) \\ & \cdot P\left(v_1 - v_2,\ v_1\left(\Delta t + \frac{d}{v_2}\right) < d\right) \end{aligned} \tag{5}$$

Control spacing d_c is considered positive if it is implemented by slowing down vehicle 2. When control actuation is implemented by slowing down vehicle 1, negative d_c is to be used in the above equation. Applying the definition of conditional probability to the above equation, the miss probability P_m becomes

$$\begin{aligned} P_m = {} & P\left(\left|\frac{v_1 - v_2}{v_2} d + v_1 \Delta t + d_c\right| < d_m,\ v_1 \geq v_2\right) \\ & + P\left(\left|\frac{v_1 - v_2}{v_2} d + v_1 \Delta t + d_c\right| < d_m,\ v_1 < v_2,\ v_1\left(\Delta t + \frac{d}{v_2}\right) \geq d\right) \\ & + P\left(\left|\frac{v_2 - v_1}{v_1} d - v_2 \Delta t + d_c\right| < d_m,\ v_1 < v_2,\ v_1\left(\Delta t + \frac{d}{v_2}\right) < d\right) \end{aligned} \tag{6}$$

Above expression for P_m can be evaluated if the probability density functions of v_1 and v_2 are known[7]. The problem of selecting the control spacing d_c such that $P_m \leq P_s$ can be carried out by an iterative procedure. This iterative procedure must simply range over possible values of d_c until a value of d_c is selected such that $P_m(d_c^*) \cong P_s$

For illustrations, let us assume v_1 and v_2 are independently normally distributed. The joint density function of v_1 and v_2 is simply the product of the normal density function $n(v_1; \overline{v}_1, \sigma_1^2)$ and $n(v_2; \overline{v}_2, \sigma_2^2)$.

$$f(v_1, v_2) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-(v_1-\overline{v}_1)^2/2\sigma_1^2 - (v_2-\overline{v}_2)^2/2\sigma_2^2} \tag{7}$$

The miss probability P_m can be found by integrating above joint density function of (v_1, v_2) over the regions specified in equation (6). We have considered such a case, and obtained some numerical value for control spacing d_c* from digital computation. Numerical values used are: $d = 500$ ft., $d_m = 50$ ft., $t = 0.5$sec., $P_s = 0.01$, $\sigma_1 = \sigma_2 = 2$ mph. All integrations were performed by truncating the probability density functions at $\pm 3\sigma$ limits. The results are tabulated in Table 1.

MEAN SPEED OF VEH. 1 (mph)

MEAN SPEED OF VEH.2 (mph)	50	55	60	65	70
50	81	28	0	0	0
55	-103	70	23	0	0
60	- 58	-108	62	17	0
65	- 16	- 65	103	56	11
70	0	- 28	-71	90	45

TABLE 1 CONTROL SPACING d_c^* FOR COMBINATIONS OF MEAN SPEEDS OF VEHICLE 1 AND 2

ON-LINE ADJUSTMENT OF CONTROL SPACING BY SEQUENTIAL DECISION THEORY

The above method of determining the merging control d_c is based on the priori knowledge of speed statistics. An on-line procedure for estimating the present value of p_m and then automatically adjusting d_c to keep $p_m \leq p_s$ can be developed based on Wald's sequential test.[8] If we conclude that $p_m > p_s$, then we will increase d_c by an amount d_c. If we conclude otherwise, we leave d_c unchanged.

To test the hypothesis $p_m \leq p_s$, we choose two values p_1 and p_u $(p_1 < p_s < p_u)$ such that if $p_m < p_1$, it is an error to reject the hypothesis and if $p_m > p_u$, it is an error of practical importance to accept the hypothesis. If $p_1 \leq p_m \leq p_u$, it is a matter of indifference as to whether the hypothesis is accepted or rejected.

Let x_i denotes the result of the merging approach by i-th vehicle pair. $x = 0$ is for a successful approach, i.e., $d_c + d_s > d_m$, whereas $x = 1$ is for the unsuccessful approach. After n merging approaches, the best estimate of miss probability $p_m{}^*$ is

$$P_m{}^* = \sum_{i=1}^{n} x_i$$

Testing of the hypothesis that $p_m < p_l$ against the alternative hypothesis $p_m > p_u$ is accomplished by observing x_i's and computing the ratio

$$\log \frac{P_u}{P_l} = n\, P_m{}^* \log \frac{P_u}{P_l} + n\,(1 - P_m{}^*)\ \frac{1-P_u}{1-P_l}$$

at each merging of a vehicle pair. The test continues as long as the ratio lies between the two previously selected number A and B. The hypothesis $p_m < p_l$ is accepted if $\log \frac{P_u}{P_l} \leq B$ and rejected if $\log \frac{P_u}{P_l} \geq A$. A and B are given approximately by

$$A = \log \frac{1-\beta}{\alpha}\ , \quad B = \log \frac{\beta}{1-\alpha}$$

where α and β represent the maximum probabilities of falsely rejecting and falsely accepting the hypothesis $p_m < p_l$ respectively.

The above sequential test procedure completely determines the acceptance number a_n and rejection number r_n. The values of a_n and r_n are given by the following equations

$$a_n = \frac{\log \frac{\beta}{1-\alpha} + n \cdot \log \frac{1-P_l}{1-P_u}}{\log \frac{P_u}{P_l} + \log \frac{1-P_l}{1-P_u}}$$

$$r_n = \frac{\log \frac{1-\beta}{\alpha} + n \log \frac{1-P_l}{1-P_u}}{\log \frac{P_u}{P_l} + \log \frac{1-P_l}{1-P_u}}$$

Therefore, to perform this sequential test, it is only necessary to compute a_n and r_n for each value of n and compare $\sum_{i=1}^{n} x_i$ with them at each merge. If $\sum_{i=1}^{n} x_i \leq a_n$, then the hypothesis $p_m \leq p_s$ is accepted and we decide not to increase d_c whereas if $\sum_{i=1}^{n} x_i \geq r_n$, then the hypothesis $p_m \leq p_s$ is rejected and d_c is increased.

In on-line operation, a table of a_n and r_n can be stored in the control computer memory so that the sequential test is simply carried out by comparisions. After each decision, computer will begin another test automatically. If no decision is reached within some upper bound of merging approaches, the computer automatically truncate the test.

CONCLUSION

An approach to merging control for existing ground traffic is presented in this paper. When the statistics of the merging vehicle speeds are known, control spacing for reducing the miss probability below a certain specified level is determined. In the case where these statistics change during the course of system operation, an on-line procedure for adjusting the control spacing based on Wald's sequential decision theory is discussed.

For the present approach to be useful in actual application, techniques to implement the control spacing determined in this paper has to be developed. This control implementation will rely mainly on the response of the human driver to the traffic control signal. Some procedures for recycling the merging vehicles in case miss does occur are also needed to ensure the operation success.

ACKNOWLEDGMENT The first author wishes to thank Mr. C. S. Koegel of Westinghouse Electric Corporation for his encouragements and the funds available for the preparation of this paper.

REFERENCES

1. G. H. Weiss and A. A. Maradudin, "Some Problems in Traffic Delay", Operations Research, Vol. 10, pp. 74-104, 1962.

2. D. R. McNeil and J. T. Smith, "A Comparison of Motorist Delays for Different Merging Strategies", Transportation Science, Vol. 3, pp. 239-254, 1969.

3. Hisashi Mine and Takeshi Mimura, "Highway Merging Problem with Acceleration Area", Transportation Science, Vol. 3, pp. 205-213, 1969.

4. D. E. Blumenfeld, "Merging From an Acceleration Lane", Transportation Science, Vol. 5, pp. 161-168, 1971.

5. M. Athans and W. S. Levine, "Optimal Vehicle Control for the Merging Problem", Report ESL-R-327, MIT Electronic Systems Laboratory, Cambridge, Mass., 1967.

6. L. P. Hajdu, K. W. Gardiner, H. Tamara and G. L. Pressman, "Design and Control Considerations for Automated Ground Transportation Systems", Proceedings of the IEEE, Vol. 56, no. 4, April 1968.

7. A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill Book Company, 1965.

8. A. Wald, Sequential Analysis, John Wiley and Sons, Inc., 1947.

THE OPTIMAL CONTROL OF MERGING AIRCRAFT - DERIVATION OF THE HYBRID AIR TRAFFIC CONTROLLER†

JOSEPH G. SCHATZ
Bell Aerospace, Division of Textron, Buffalo, New York

THOMAS E. STERN
Columbia University, New York, New York

ABSTRACT

This paper considers the development of an exactly realizable and easily implementable automatic controller to merge aircraft in a safe and feasible manner. In order to obtain such a controller, the merging problem is formulated as a *finite-time,* quadratic, optimal control problem of a linear system with *state and control constraints*. By employing a non-autonomous control structure, a controller known as the Hybrid Air Traffic Controller, is developed without use of numerical constrained optimization techniques.

I. INTRODUCTION

In recent years the Federal Aviation Administration has initiated a program to automate its existing manual air traffic handling procedures [1, 2]. The purpose of this program is to increase the efficiency of the manual controller by automating many of the routine procedures now performed. One such procedure, whose automation is expected to increase controller efficiency by 2.6 to 2.9 times, is known as aircraft merging [3]. This task requires that N aircraft distributed among K feeder routes be merged into one or more final routes such that the following requirements are satisfied:

a. At the completion of the procedure all aircraft are traveling at a *prescribed velocity* and are separated by a *prescribed separation.*
b. The procedure is performed *safely* (all aircraft must be separated by at least a minimum distance) and *feasibly* (no aircraft is required to perform maneuvers which exceed its performance limitations).
c. The procedure must be completed in a *finite time* T.

This paper considers the question of obtaining an exactly realizable and easily implementable automatic controller for merging which satisfies requirements (a) through (c). The problem of obtaining such a controller is formulated as a problem of the *optimal control* of a *linear system* which must be brought to the origin in a *finite time* while minimizing a *quadratic cost* functional. The safety and feasibility of the procedure are insured in this formulation by placing *constraints* on the allowable states and controls.

Previous investigations of merging in papers by Athans [4] and Porter [5] have obtained automatic controllers for the merging procedure by formulating it as an optimal linear regulator problem [6]. Therefore, the solutions obtained satisfy only requirement (a), and do not in general represent safe, feasible and finite-time merging.

---

† This paper is based on research submitted in partial fulfillment of the requirements for the Degree of Doctor of Engineering Science in the School of Engineering and Applied Science, Columbia University.

This research was supported in part by the National Aeronautics and Space Administration under Grant NGL-33-008-090 and by the National Science Foundation under Grant GK-2283.

II. MODEL

For the purpose of this paper it will be assumed that there is a single final route and that both final and feeder routes are linear and radiate from a single point known as the Merging Point (MP). This situation is illustrated in Fig. 1. In addition, it will also be assumed that the prescribed velocity and separation referred to in requirement (a) are constants and not aircraft dependent. These constants will be referred to as the *string velocity* V and the *standard separation* Δ respectively.

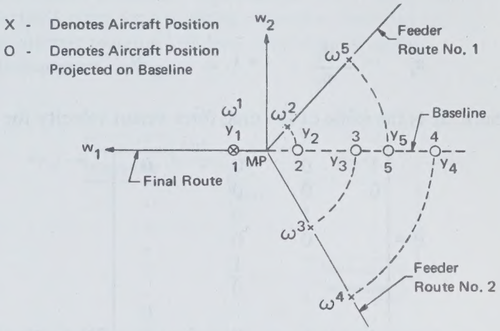

X - Denotes Aircraft Position
O - Denotes Aircraft Position Projected on Baseline
w_2
w_1
Feeder Route No. 1
Feeder Route No. 2
Baseline
Final Route
MP
ω^1
ω^2
ω^3
ω^4
ω^5
y_1
y_2
y_3
y_4
y_5
1
2
3
4
5

Fig. 1. Example of Reduction of Merging Problem to String Problem

As demonstrated in [5] the merging problem may now be reduced to a *string problem* by a circular projection (using MP as the center) of the location of the aircraft in the feeder routes onto the horizontal extension of the final route (referred to as the baseline). This technique is illustrated for four aircraft in Fig. 1.

Once the aircraft merging problem has been reduced to the equivalent aircraft string problem, the resulting system may be modeled by the following set of linear perturbation state equations.

$$\dot{x}(t) = A\,x(t) + B\,u(t)\ ;\ x(t_o) = x_o \tag{1}$$

Where the state vector x (t) and the control vector u (t) are defined as:

$$\begin{aligned} x_{2j}(t) &= y_j(t) - y_{j+1}(t) - \Delta\ ;\ j = 1, 2, \ldots, N\text{-}1 \\ x_{2j-1}(t) &= v_i(t) - V \\ u_i(t) &= (f_i(t) - G_i)/m_i\ ;\ i = 1, 2, \ldots\ldots, N \end{aligned} \tag{2}$$

and for the i^{th} aircraft in the merging sequence:

$y_i(t)$ is its position on the baseline at time t,
$v_i(t)$ is its velocity at time t,
$f_i(t)$ is the engine thrust applied at time t,
G_i is the drag on the aircraft when traveling at V,
m_i is its mass, which is assumed constant.

These defintions are illustrated in Fig. 2. In addition, matrices A and B are given as follows:

$$A = \begin{bmatrix} -a_1 & 0 & & & & \\ 1 & 0 & -1 & & \bigcirc & \\ & 0 & -a_2 & 0 & & \\ & & 1 & \ddots & \ddots & \\ & \bigcirc & & \ddots & \ddots & -1 \\ & & & & \ddots & -a_N \\ & & & & 0 & \end{bmatrix}_{2N-1,\ 2N-1}$$

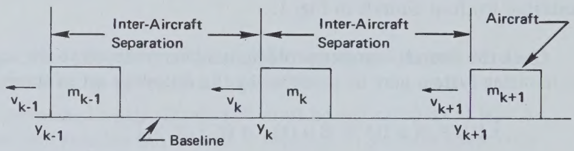
$$a_i = \frac{\alpha_i}{m_i}, \quad i = 1, 2, \ldots, N$$

where α_i is the slope of the drag force versus velocity for aircraft i when $v_i = V$.

$$B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & \vdots \\ \vdots & 1 & 0 & & \vdots \\ \vdots & 0 & 0 & & \vdots \\ \vdots & \vdots & 1 & & \vdots \\ \vdots & \vdots & 0 & & \vdots \\ \vdots & \vdots & \vdots & & 0 \\ 0 & 0 \cdots & 0 & 0 & 1 \end{bmatrix}_{2N-1,\ N}$$

This model is derived in [7, 8].

Fig. 2. String of Aircraft



The questions of obtaining an automatic controller for merging may now be formulated as a problem of the optimal control for the linear system (1) which must be brought to the origin in a finite time, i.e.:

$$x(T) = 0 \tag{3}$$

such that the following quadratic cost functional is minimized

$$J = \frac{1}{2} \int_{t_o}^{T} \left| x^t Qxt\, u^t R u \right| dt \tag{4}$$

where Q and R are appropriately dimensioned, positive-definite, diagonal weighting matrices. The safety and feasibility of the merging procedure are insured in this formulation by placing the following constraints on both the states and controls.

$$\alpha_i^{x} \leqslant x_i(t) \leqslant \beta_i^{\ x} \ ; \ i = 1, 2, \ldots, 2N\text{-}1 \tag{5}$$

$$\alpha_j^{x} \leqslant u_j(t) \leqslant \beta_j^{u} \ ; \ j = 1, 2, \ldots, N$$

where the bounding constants are related to the type of aircraft being merged.

III. DEVELOPMENT OF CONTROLLER

The technique used to solve the constrained optimization problem (1-5) is first to solve the unconstrained problem (1-4). Sufficient conditions are then developed, which, if satisfied, will guarantee that the solution of this latter problem will meet the safety and feasibility constraints (5).

It is well known that the unconstrained fixed-time optimization problem (1-4) leads to an autonomous feedback control system which cannot be exactly implemented [9]. This situation can be circumvented by making use of a partially open-loop partially closed-loop control structure [10]. This structure, as applied to the aircraft merging problem, is referred to as the *Hybrid Air Traffic Controller* (HAC) and is illustrated in Fig. 3.

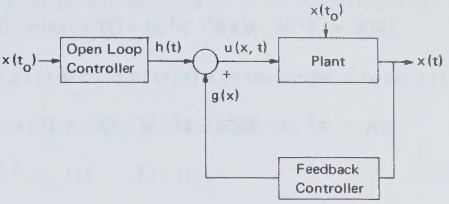


Fig. 3. Hybrid Controller (HAC)

The Hamiltonian for the minimization problem defined by equations (1-4) is independent of controller structure and is given by

$$H = \frac{1}{2}\left|x^t(t)\,Qx(t) + u^t(t)\,Ru(t)\right| + p^t(t)\left|Ax(t) + Bu(t)\right| \tag{6}$$

where p (t) is the 2N-1 dimensional co-state vector.

Application of the Minimum Principle of Pontryagin to (6) gives the following necessary conditions which x (t) and u(t) must satisfy in order that the cost functional (4) be minimized

$$\begin{bmatrix} \dot{x}(t) \\ \dot{p}(t) \end{bmatrix} = \left[\begin{array}{c|c} A & -BR^{-1}B^t \\ \hline -Q & -A^t \end{array}\right] \begin{bmatrix} x(t) \\ \hline p(t) \end{bmatrix} = \bar{A} \begin{bmatrix} x(t) \\ \hline p(t) \end{bmatrix} \tag{7}$$

$$u(t) = -R^{-1}B^t\,p(t) \tag{8}$$

where the 4N-2 by 4N-2 matrix A has the obvious definition.

The boundary conditions in (7) are

$$x(T) = 0 \;;\; x(t_o) = x_o. \tag{9}$$

The mathematical formaulation of the HAC control structure as indicated in Fig. 3 is

$$u(x, t) = g(x) + h(t) \tag{10}$$

where:

g (x) is a N vector dependent only upon state feedback, and
h (t) is a N vector representing an open loop control.

Now assume that the co-state vector is given by

$$p(t) = p(x,t) = Mx(t) + f(t) \tag{11}$$

where:

M is a constant 2N-1 by 2N-1 *feedback gain matrix*, and
f (t) is a 2N-1 dimensional vector.

If (11) is now differentiated, the following is obtained

$$\dot{p}(x,t) = M\dot{x}(t) + \dot{f}(t) \, . \tag{12}$$

Substitution of (7) and (11) into (12) yields

$$[MA + A^t M - MBR^{-1} B^t M + Q]\, x(t) = -\dot{f}(t) - (A^t - MBR^{-1} B^t)\, f(t) \, . \tag{13}$$

If f (t) is to be independent of x (t),as implied in (11), then (13) implies that

$$MA + A^t M - MBR^{-1} B^t M + Q = 0 \tag{14}$$

and

$$\dot{f}(t) + (A^t - MBR^{-1} B^t)\, f(t) = 0. \tag{15}$$

Since the system equation (1) can easily be shown to describe a completely controllable system,the steady-state Riccati equation (14) has a unique, positive definite, symmetric solution [6]. All that remains, then, is to obtain the initial condition for (15).

If (11) is substituted into (8) the control is given by

$$u(x,t) = -R^{-1} B^t Mx(t) - R^{-1} B^t f(t) \, . \tag{16}$$

Therefore, the system equation (1) becomes

$$\dot{x}(t) = (A - BR^{-1} B^t M)\, x(t) - BR^{-1} B^t f(t) \tag{17}$$

The solution of (17) is given by

$$x(t) = \bar{\Phi}(t, t_o)\, x(t_o) - \int_{t_o}^{t} \bar{\Phi}(t,\alpha)\, BR^{-1} B^t f(\alpha)\, d\alpha \tag{18}$$

where $\bar{\Phi}(t, t_o)$ is the transition matrix for the homogeneous differential equation

$$\dot{x}(t) = (A - BR^{-1} B^t M)\, x(t) \tag{19}$$

Comparing (15) and (19), it is observed that they describe the motion of adjoint systems. Therefore,

$$f(t) = \bar{\Psi}(t, t_o)\, f_o \tag{20}$$

where

$$\bar{\Psi}(t, t_o) = \bar{\Phi}^t (t_o, t).$$

Equation (18) can then be rewritten as

$$x(t) = \bar{\Phi}(t, t_o)\ x(t_o) - \int_{t_o}^{t} \bar{\Phi}(t, \alpha) BR^{-1} B^t\ \bar{\Psi}(\alpha, t_o)\, d\alpha f_o. \tag{21}$$

When the terminal condition x (T) = 0 is introduced into (21), the following expression is obtained for f_o.

$$x(t_o) = \mathcal{G}(T) f_o \tag{22}$$

where

$$\mathcal{G}(t) \triangleq \int_{t_o}^{t} \bar{\Phi}(t_o, \alpha) BR^{-1} B^t\ \bar{\Phi}^t(t_o, \alpha)\, d\alpha .$$

However, since the system described by (1) is completely controllable, it can be shown [13] that $\mathcal{G}(t)$ is nonsingular for $t > t_o$. Therefore, using (20)

$$f(t) = \bar{\Psi}(t, t_o) \mathcal{G}^{-1}(T)\ x(t_o).$$

The control law, u (x, t), may now be rewritten in the following form,

$$u(x, t) = -R^{-1} B^t [M x(t) + \bar{\Phi}^t(t_o, t) \mathcal{G}^{-1}(T) x(t_o)]. \tag{23}$$

comparing (23) with (10), the closed loop control is

$$g(x) = -R^{-1} B^t M x(t) \tag{24}$$

and the open loop control is

$$h(t) = -R^{-1} B^t\ \bar{\Phi}^t(t_o, t) \mathcal{G}^{-1}(T) x(t_o). \tag{25}$$

If (28) is substituted into (21) the following expression is obtained for the state of the merging process

$$x(t) = \Phi^x(t, t_o)\ x(t_o) \tag{26}$$

where

$$\Phi^x(t, t_o) = \bar{\Phi}(t, t_o) [I - \mathcal{G}(t) \mathcal{G}^{-1}(T)].$$

If (26) is now substituted into (23) the control at time t may be written as

$$u(t) = \Phi^u(t, t_o)\ x(t_o) \tag{27}$$

where

$$\Phi^u(t, t_o) = -R^{-1} B^t \left\{ M\ \bar{\Phi}(t, t_o) [I - \mathcal{G}(t) \mathcal{G}^{-1}(T)] + \bar{\Phi}^t(t_o, t) \mathcal{G}^{-1}(T) \right\}.$$

Equations (26 - 27) represent closed form expressions for the merging process optimal controls and trajectories. It is these expressions which will form the basis for the derivation of the sufficient conditions which insure safety and feasibility.

It has been shown in [13] that (24 - 27) may be easily implemented by making use of the relationship between the eigenvalues and eigenvectors of the $\bar{A}$ matrix and the solution of the steady-state Riccati equation [11 - 12]. In addition, since it is possible to determine *closed* *form* expressions for these eigenvalues and eigenvectors [13], equations (24 - 27) may be implemented in *real-time* on a digital computer.

IV. REGIONS OF FEASIBLE AND SAFE MERGING

The merging problem formulation which led to the development of the HAC ignored the safety and feasibility requirements (5). This section will define regions in the state space in which the initial state $x(t_o)$ must be located if the controlled aircraft are to be merged in a feasible and safe manner.

The safe and feasible merging of aircraft over the interval $I_1 = \{t \mid t_o \leqslant t \leqslant T\}$ will be defined as follows.

Definition 1:

A particular set of merging controls and states will be referred to as feasible if:

(a) the aircraft control deviations satisfy the following inequality

$$\alpha_i^u \leqslant u_i(t) \leqslant \beta_i^u \quad ; \quad i = 1,2,\ldots,N, \quad t \in I_1 \tag{28}$$

(b) the aircraft velocity deviations satisfy the following inequality

$$\alpha_i^x \leqslant x_i(t) \leqslant \beta_i^x \quad ; \quad i = 1,3,\ldots,2N\text{-}1, \quad t \in I_1 \tag{29}$$

where the constants α_i^x, α_i^u, β_i^x, β_i^u are determined by the capabilities of the merging aircraft.

Definition 2:

A particular set of merging states will be referred to as safe if the aircraft deviations from the standard separation satisfy the following inequalities:

$$\begin{aligned} -\epsilon \leqslant x_i(t) \quad &; \quad i = 2,4,\ldots,2N\text{-}2 \\ &; \quad t \in I_1 \end{aligned} \tag{30}$$

where $0 \leqslant \epsilon < \Delta$ such that $\Delta - \epsilon \geqslant d_{min}$ and d_{min} is the minimum allowable separation between aircraft.

In order to determine feasibility and safety the appropriate elements of the $\Phi^x(t, t_o)$ and $\Phi^u(t, t_o)$ matrices defined in (26-27) will be bounded from above and below. These bounds will then be used to determine a set of sufficient conditions on $x(t_o)$ which ensure that (28-29 and 30) are satisfied.

Let $n_{i,j}$ and $N_{i,j}$ be the minimum and maximum values of the (i, j) element of $\Phi^x(t, t_o)$ over the interval I_1, and let $m_{i,j}$ and $M_{i,j}$ be the minimum and maximum values of the (i,j) element of $\Phi^u(t, t_o)$ over the interval I_1. When those bounds are substituted into (26-27), it is found that the merging states are feasible and safe if the elements of the initial state vector satisfy the following unequalities.

(a) Feasibility

$$\alpha^{x}_{i} \leqslant \sum_{j=1}^{2N-1} x_j(t_o)\ \delta^{x}_{i,j}\ (x_j(t_o))\ ; \quad \beta^{x}_{i} \geqslant \sum_{j=1}^{2N-1} x_j(t_o)\ \delta^{x}_{i,j}\ (-x_j(t_o))$$

$$i = 1,3, \ldots, \ 2N\text{-}1 \tag{31}$$

$$\alpha^{u}_{i} \leqslant \sum_{j=1}^{2N-1} x_j(t_o)\ \delta^{x}_{i,j}\ (x_j(t_o))\ ; \quad \beta^{u}_{i} \geqslant \sum_{j=1}^{2N-1} x_j(t_o)\ \delta^{u}_{i,j}\ (-x_j(t_o))$$

$$i = 1,2, \ldots, N. \tag{32}$$

(b) Safety

$$\sum_{j=1}^{2N-1} x_j(t_o)\ \delta^{x}_{i,j}\ (x_j(t_o)) \geqslant -\epsilon; \quad i = 2,4, \ldots, 2N\text{-}2 \tag{33}$$

where

$$\delta^{x}_{i,j}\ (x_j(t_o)) = \begin{cases} n_{i,j} & \text{if } x_j(t_o) \geqslant 0 \\ N_{i,j} & \text{if } x_j(t_o) < 0 \end{cases}$$

$$\delta^{u}_{i,j}\ (x_j(t_o)) = \begin{cases} m_{i,j} & \text{if } x_j(t_o) \geqslant 0 \\ M_{i,j} & \text{if } x_j(t_o) < 0 \end{cases}$$

The 4N relations (31-32) and the N-1 relations (33) define two sets of hyperplanes in the state space. The regions bounded by these hyperplanes will be referred to as the Region of Feasible Merging (RFM) and the Region of Safe Merging (RSM), i.e.,

$$\text{RFM} = \{x(t_o) \mid x(t_o) \text{ satisfies equations (31-32)}\} \tag{34}$$

$$\text{RSM} = \{x(t_o) \mid x(t_o) \text{ satisfies (33)}\} \tag{35}$$

Thus, merging will be feasible if $x(t_o) \epsilon$ RFM and safe if $x(t_o) \epsilon$ RSM. It should be noted that these sufficient conditions for the constrained optimization problem (1-5) avoid use of the classical numerical techniques of constrained optimization.

An illustration of the RFM (RSM) in two dimensions is given in Fig. 4 (Fig. 5). If a given initial state is an element of the RFM (RSM) then the resultant optimal merging maneuvers implemented by the HAC will be feasible (safe). Such a state is represented by point A in Fig. 4 (Fig. 5). If a given state is not an element of the RFM (RSM) the resultant merging maneuver may not be feasible (safe). Since an Air Traffic Control System is usually designed in a highly conservative manner, an initial state which is not an element of the RFM (RSM) will be referred to as an infeasible (unsafe) initial state. Such a state is represented by point B in Fig. 4 (Fig. 5). In addition, there is always the possibility that the RFM (RSM) or the RFM $\cap$ RSM may be empty.

This condition would indicate that the HAC would require redesign making use of different Q and R matrices and possibly a new (larger) final time.

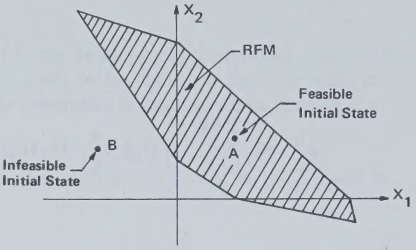


Fig. 4. Illustration of Region of Feasible Merging

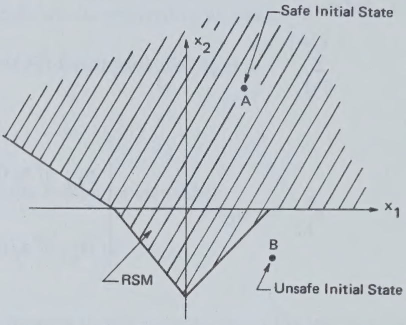


Fig. 5. Illustration of Region of Safe Merging

The required bounds for the elements of $\Phi^x(t, t_o)$ and $\Phi^u(t, t_o)$ may be easily calculated by numerically searching, over the interval I_1, the closed-form expressions developed for these matrices. In practice, these bounds should be computed off-line for each merging sequence and then stored as part of the HAC.

V. CONCLUSIONS

In order to merge aircraft in a finite time,a partially open-loop partially closed-loop control structure has been employed. This structure, referred to as the Hybrid Air Traffic Controller (HAC),has resulted in an easily implemented automatic controller to merge aircraft. In addition, since it has been possible to develop closed form expressions for the optimal merging controls and trajectories, sufficient conditions have been developed to ensure safe and feasible merging. This technique avoids use of numerical techniques used for constrained state and control optimization problems.

REFERENCES

[1] U.S. Task Force on Air Traffic Control. *Report of Task Force on Air Traffic Control, Project Beacon.* Washington, D.C.: Federal Aviation Administration, 1961.

[2] Department of Transportation Economics Division Office of Policy Development. *Aviation Forecasts 1969-1980.* Washington, D.C.: Federal Aviation Administration, Jan. 1969.

[3] Department of Transportation Control Advisory Committee. *Report of Department of Transportation Control Advisory Committee.* Vol. 2, Washington, D.C.: Government Printing Office, Dec. 1964.

[4] Athans, M., "A Unified Approach to the Vehicle-Merging Problem." *Transportation Research,* Pergamon Press, Vol. 3 (1969), 123-133.

[5] Porter, L., "On the Optimal Scheduling and Holding Strategies for the Air Traffic Control Problem." Report ESL-R-401, M.I.T. Electronic Systems Laboratory, Cambridge, Mass. Sept. 1969.

[6] Athans, M., and Falb, P., Optimal Control: *An Introduction to the Theory and Application.* New York: McGraw Hill, 1966.

[7] Levine, W. and Athans, M,, "On the Optimal Error Regulation of a String of Moving Vehicles." *IEEE Transactions on Automatic Control,* Vol. AC-11, No. 3 (July 1966), 355-361.

[8] Athans, M., Levine, W., and Leves, A., "A System for the Optimal and Suboptimal Position and Velocity Control for a String of High Speed Vehicles." Report ESL-P-314, M.I.T. Electronic Systems Laboratory, Cambridge, Mass., July 1967.

[9] Friedland, B., "On the Solution of the Riccati Equation in Optimization Problems," *IEEE Transactions on Automatic Control,* Vol. AC-12, No. 3 (June 1967), 303-304.

[10] Rekasius, Z., "An Alternate Approach to the Fixed Terminal Time Problem." *IEEE Transactions on Automatic Control,* Vol. AC-9, No. 3 (July 1964), 290-292.

[11] O'Donnell, J., "Asymptotic Solution of the Matrix Riccati Equation of Optimal Control." *Proceedings of the Fourth Allerton Conference on Circuit and System Theory,* (1966).

[12] Potter, J., "Matrix Quadratic Solutions." *Journal of the SIAM on Applied Mathematics,* Vol. 4, No. 3 (May 1966).

[13] Schatz, J.G., "The Optimal Control of Merging Aircraft." PhD thesis, Columbia University Department of Electrical Engineering, 1971.

A STOCHASTIC NETWORK MODEL FOR AIR CARGO TERMINALS

CHAO LEE* AND LAWRENCE P. McNAMEE**
*Litton Data Systems Division, Woodland Hills, California
**University of California, Computer Science Department
Los Angeles, California

ABSTRACT

Bottlenecks of the material handling operation of an air cargo terminal are investigated by the stochastic network method of GERT IIIQ. The relationship of the GERTS IIIQ network model to construction specification is discussed. Operational bottlenecks are identified and corrected through modification of building constraints.

INTRODUCTION

In the past, design of air cargo terminals was usually based upon an expected life-time of twenty to twenty five years including approximately three to five years construction lead-time. When such designs proved inadequate, makeshift measures such as replacing material handling equipment with faster machinery, or expanding the terminal building itself proved to be expedient and quite acceptable.

In recent years, however, with the introduction of jumbo jets, traditional policies had to be scrapped and new radically different concepts examined. No longer can makeshift modifications suffice.

Two design influencing factors are apparent: the volume-tonnage, which will increase by an order of magnitude over the next five years, and the number of individual pieces of cargo which will increase accordingly. It is obvious then that reasonable anticipatory precautions in construction concepts such as can be obtained through simulation studies must be included in the early stages of the design procedure.

The objectives of this paper are: (1) to develop a simulation model of an air cargo facility, (2) to identify air terminal operational bottlenecks, and (3) to present construction strategies to correct or compensate for such bottlenecks.

The paper is divided into five sections. In the next section the general capabilities of the GERTS IIIQ, a stochastic event simulation program, which is used in the simulation, is described. This is followed by a discussion of the functional characteristics of a typical air cargo facility. The simulation model and studies are given in Section 4. The last section summarizes results of the paper and discusses future areas of research.

THE GERTS CONCEPT

GERTS IIIQ (Graphical Evaluation and Review Technique), as defined by Pritsker [1], is a procedure for formulating and evaluating systems using a graphical approach. The GERTS IIIQ approach utilizes the following steps: (1) Convert a qualitative description of a system to a stochastic network model; (2) Collect data necessary to describe the function ascribed to the branches of the network; (3) Conduct simulation studies to determine performance measurements; (4) Make inferences concerning the system under investigation from studies made in 3 above.

GERTS IIIQ Capabilities: GERTS IIIQ is particularly suited for engineers and nonprogrammers who want to conduct simulation studies from a graphical or network point of view, and not be concerned with detail programming of events and activities such as is done with Simscript and GPSS. With GERTS IIIQ, problem insight is immediately available.

Node Characteristics: GERTSIIIQ networks consist of nodes and directed branches. Each node represents an event and is depicted by one of the symbols shown in Table 2.1 according to the function to be performed. A node is deterministic if the output branches (activities) are all scheduled, i.e. the probability is one that all branches emanating from the node are taken. A node is probabilistic if only one branch (activity) emanating from the node is scheduled. The branch selected is determined by a random number generator; the sum of all probabilities of branches emanating from the probabilistic node is one. A semicircle on the output side of the node denotes a deterministic node; a triangle denotes a probabilistic node.

For the basic (source, sink, and event) nodes, the A parameter specifies the number of times activities (branches) incident to a node must be completed before the node (event) is realized. The B parameter specifies the number of times the activities incident to the node must be completed after the node is realized the first time to realize the node again and to reschedule activities. The N parameter is a number assigned to each node; all nodes must be assigned numbers and all numbers must be unique.

The parameters on the input side of the queue nodes have a completely different meaning. The C parameter specifies the initial number of items in the queue; the D parameters the maximum number of items allowed in the queue. When an activity incident to the queue node is completed, one of two possible actions can take place; either the activity following the node is initiated, or the number in the queue is increased by one.

The activity emanating from a queue node represents a service activity. With the present version of the program, only one service activity can be associated with a queue node. (Note that a probabilistic output is allowed since only one activity will be serviced at any one time.)

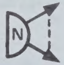
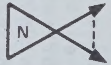
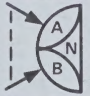
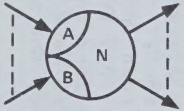
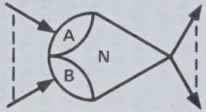
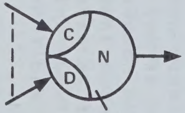
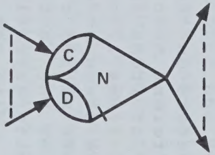
Whenever the number of items in the queue exceeds the maximum number allowed (parameter D), the activity can balk from the system as if it never occurred in the first place, or it can balk to another node which must be a basic node, (and not another queue node), and immediately the activities emanating from it are scheduled.

Branch Characteristics: The branches of GERTSIIIQ networks represent activities and/or information transfer. Associated with each branch are the following characteristics: (1) The probability that the activity will be initiated given the node it emanates from is realized; (2) The time to perform the activity once it is started. The time function can be distributed according to one of the available types: (a) Constant, (b) Normal, (c) Uniform, (d) Erlang, (e) Lognormal, (f) Poisson, (g) Beta, (h) Gamma, and (i) Beta fitted to three parameters as in PERT; (3) A Counter that records the number of times a group of activities is exercised before a node is realized; (4) An activity number which permits network modifications.

Counters can be assigned to a branch or a group of branches. The number of counts recorded is always referenced to a particular node. Network modifications are possible by assigning an

TABLE 2.1

GERT IIIQ NODE CAPABILITIES

PROGRAMS	NODE FUNCTION	DETERMINISTIC SYMBOL	PROBABILISTIC SYMBOL	PARAMETERS
Basic to Gert III	Source			A = No. of times activities incident to the node must be realized for node to be realized B = No. of times activities incident to the node must be realized after the first time the node is realized to repeat N = Node number
	Sink[1]		Not Applicable	
	Event			
Added to Form Gert IIIQ	Queue			C = Initial number in the queue D = Maximum number allowed in queue N = Node number

Note: [1] The sink node can not really be classified in this manner since it can be an event node with a sink code assigned to it. The symbol shown is the most common form used and is included here as a matter of convenience.

activity number to a branch, and when the branch is activated, modification can take place.

Output Capabilities: GERTSIIIQ simulates a given network a user-specified number of times, collects statistics, and prints the results. The statistics printed are: (1) The probability a node is realized. (2) The mean, standard deviation, minimum and maximum times, associated with a user specified node, and for how many observations involved. One of the following types of time statistics are allowed: (a) the time of first realization of a node, (b) the time of all realizations of a node, (c) the time between realizations of a node, (d) the time interval required to go between two nodes in the network, (e) the time delay from the first activity completion of the node until the node is realized again. (3) A histogram of the time to realize a node as defined in 2 above. (4) The mean, standard deviation, minimum and maximum number of counts to realize a node.

Unless specified by the user first time statistics are automatically collected on all queue and sink nodes. Simulation traces of runs can also be obtained to check detail operation of the network.

Program Operation: The GERTSIIIQ program performs a simulation of the network by advancing time from event to event. The start nodes are realized at time equal to zero and schedules all activities from them. Each event is realized just as described previously, until a specified number of sink nodes are realized. This completes a simulation run. This procedure is repeated until all the desired simulation runs are performed, and the statistics printed.

After this is accomplished inferences on system performances can be made, or simulations rerun with a new set of parameters or conditions. In the air cargo simulation studies that follow, inferences abstracted in terms of construction parameters are described.

REQUIREMENTS FOR AIR CARGO FACILITIES

In order to model air cargo facilities it is necessary to identify the appropriate functional operations in the terminal, including the pertinent parameters, and to establish relationships between these operations.

The major air freight terminal operations are: (1) Receive and ship freight by land carriers, such as trucks and trains, and by jumbo jets such as the C-5A and C-141. (2) Inspect, document, and sort cargo according to destination. (3) Transfer cargo within the terminal for processing. (4) Provide storage for aggregation of cargo for future shipments. (5) Breakdown and buildup of cargo into manageable units for reshipping. (6) Handle special and/or high priority freight.

To carry out all of these functions it is necessary to provide a coordinated operation involving manpower, both moveable and stationary material handling equipment, and control. The important operational parameters are: (1) The Tonnage-volume of the cargo to be handled at each station, (2) The speed at which each operation is conducted, and (3) The control or branching between stations.

The basic considerations in dealing with the volume of cargo at stations and with the speed of processing it are: (1) The Method of Unitization: The primary methods of aggregating cargo for shipment are by pallets and containers. Pallets offer a better time advantage, but do not provide sturdiness and protection from the elements. Pallets require less storage than containers when empty. (2) Speed and Width of Conveyors: The amount of cargo conveyors can handle is related to the line speed and width. They can sort large volumes of cargo almost immediately after unloading at the truck dock according to destinations. (3) The Number, Speed, and Payload of Moveable Material Handling Equipment: Forklifts and K-loaders probably constitute the most flexible cargo handling machines used in the terminal to move cargo between stations, and to load and unload cargo at docks. Their rate of cargo transfer is dictated by the average speed of operating between stations, and the sizes of payloads they can carry. (4) Speed and Space Required by Heavy Moving Equipment: The working space required by cranes, in terms of tonnage and placement in the terminal operation can easily be computed. (5) Storing and Retrieval Speeds: Average times for storing and retrieving cargo are determined by the working space and the distance between the storage area and the stations to or from which it is moved. (6) Areas Provided for Pallet Buildup-Breakdown: A large working space is needed to buildup and/or breakdown cargo since both pallets and containers should be handled. Average times for this operation is determined by the proximity of cargo storage area to the buildup/breakdown area and the type of equipment feeding into it. This is one of the most crucial operations in the entire facility.

All or most of these operations are common to air cargo terminals. Recently Dortech introduced a new approach to air cargo handling which utilizes towline carts for station to station transfer within the terminal building, and stacker bins for aggregation of cargo for future shipments.

The towline carts are 8 ft. long, 4 ft. wide, and 5 ft. high with an average load capacity of 1200 lbs. Transportation of these carts is by means of a towline built into the terminal floor. The destination is determined by selecting a particular towpin which is programmed for transportation to a specific station. The empty cars are returned in a similar manner to the loading docks or to other stations.

Cargo, along with its towline cart, can be randomly stored in stacker bins by cranes which can move both laterally and vertically. The stacker bins are located in a structure 150 ft. wide and 40 ft. high. Cargo can easily be retrieved upon command.

The towline operation requires less equipment space to transfer cargo between stations and sort cargo than a conveyor system. However it must be kept in mind that the reduction in space is reflected by the addition of the stacker bin storage area.

The next section describes the simulation of a Dortech type air cargo facility.

AN AIR CARGO SIMULATION MODEL

A GERTS IIIQ simulation model of a Dortech type facility is shown in Figure 4.1. Arriving truck freight is simulated by source node 2 and event node 3; arriving airplanes by source 50 and event node 51. Truck and plane departures are represented by sink nodes 45 and 90, respectively. Trucks are unloaded according to one of the following cargo types: multipallets, cartable, noncartable/oversize, pallet, and priority/special handling. Both C-5A and C-141 aircraft are accommodated. The act-

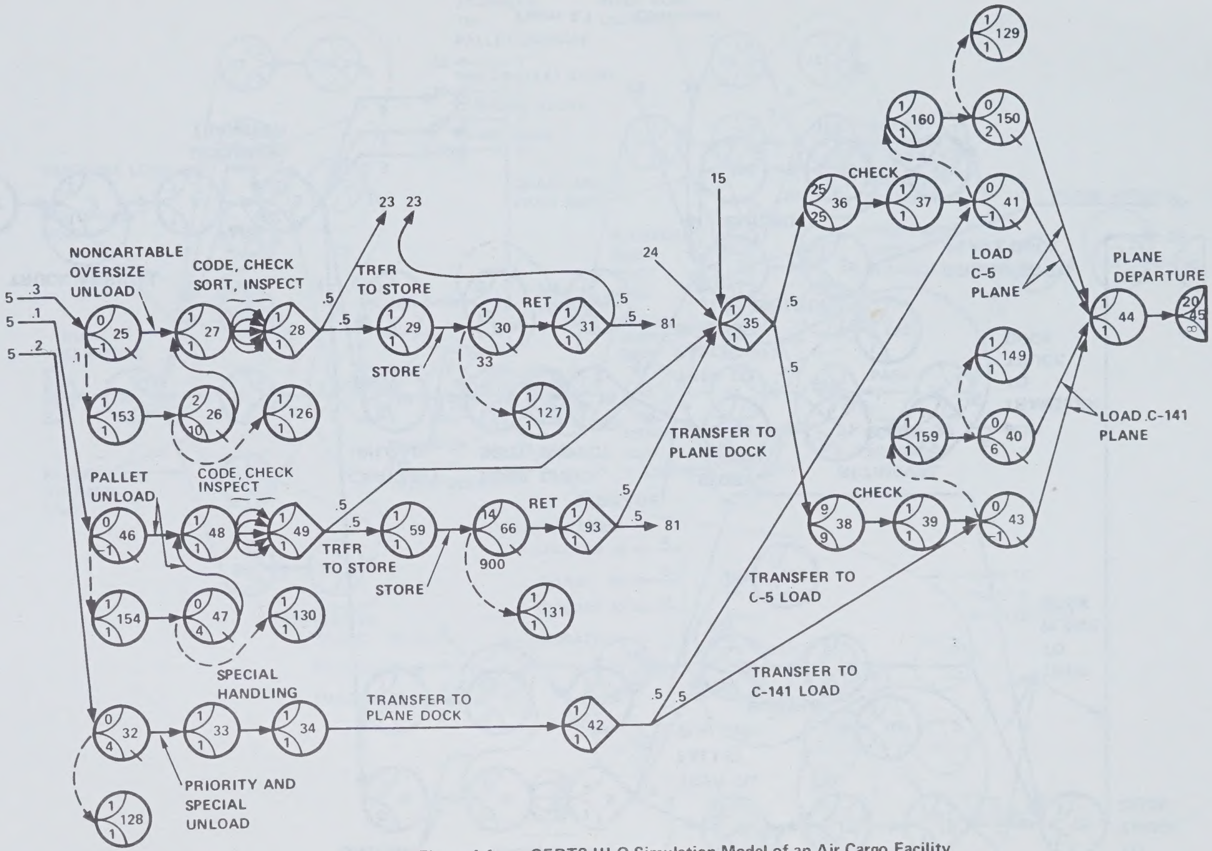

NONCARTABLE OVERSIZE UNLOAD
CODE, CHECK SORT, INSPECT
TRFR TO STORE
STORE
RET
23 23
24
15
TRANSFER TO PLANE DOCK
CHECK
LOAD C-5 PLANE
PLANE DEPARTURE
LOAD C-141 PLANE
PALLET UNLOAD
CODE, CHECK INSPECT
TRFR TO STORE
STORE
RET
900
33
81
81
TRANSFER TO C-5 LOAD
TRANSFER TO C-141 LOAD
SPECIAL HANDLING
TRANSFER TO PLANE DOCK
PRIORITY AND SPECIAL UNLOAD
25
27
28
29
30
31
35
36
37
41
150
160
129
44
45
153
26
126
127
46
48
49
59
66
93
131
154
47
130
32
33
34
42
38
39
43
40
159
149
128

Figure 4.1. GERTS III Q Simulation Model of an Air Cargo Facility.

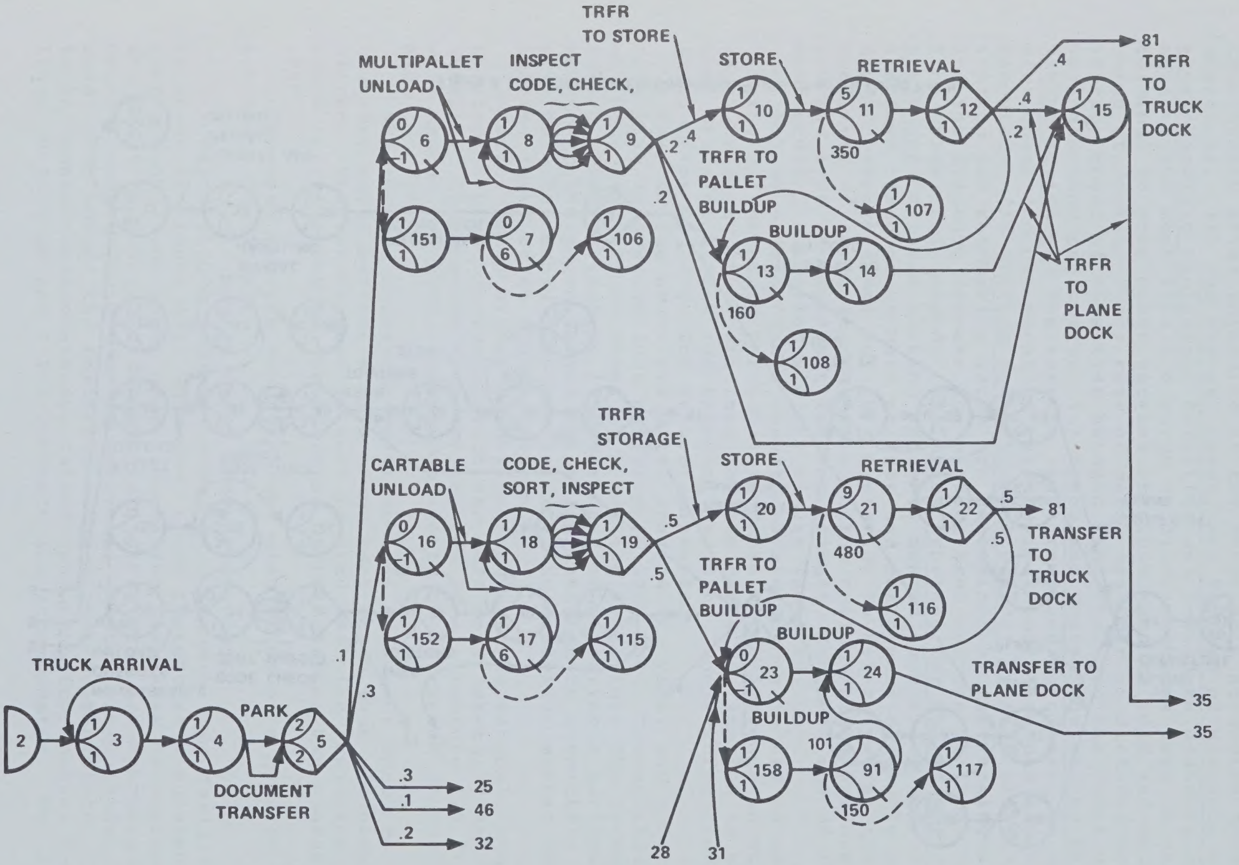
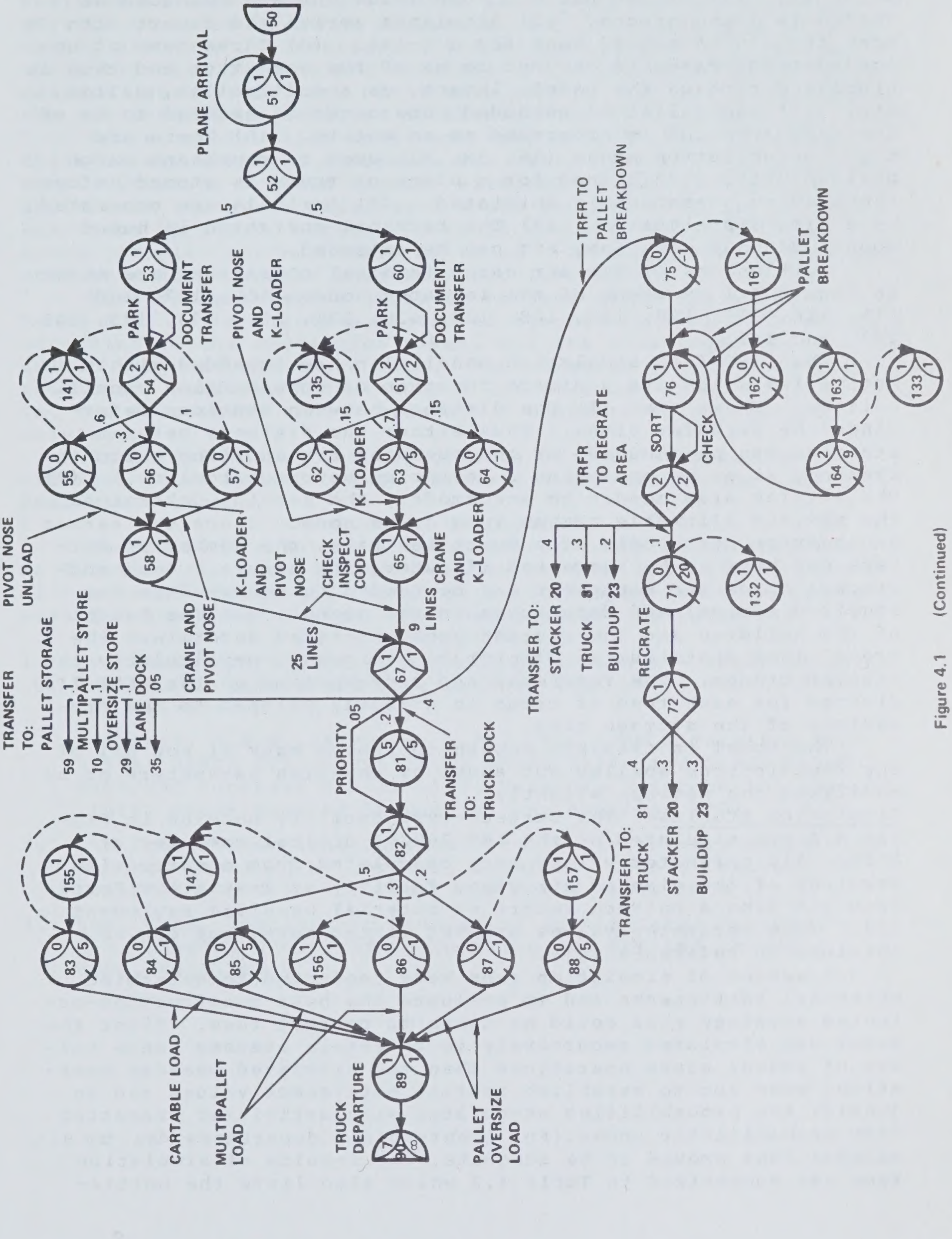

TRUCK ARRIVAL
PARK
DOCUMENT TRANSFER
MULTIPALLET UNLOAD
INSPECT CODE, CHECK,
TRFR TO STORE
STORE
RETRIEVAL
TRFR TO PALLET BUILDUP
BUILDUP
81
TRFR TO TRUCK DOCK
TRFR TO PLANE DOCK
CARTABLE UNLOAD
CODE, CHECK, SORT, INSPECT
TRFR STORAGE
STORE
RETRIEVAL
81
TRANSFER TO TRUCK DOCK
TRFR TO PALLET BUILDUP
BUILDUP
BUILDUP
TRANSFER TO PLANE DOCK
35
35
25
46
32
28
31
350
160
480
101
150
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
91
106
107
108
115
116
117
151
152
158

Figure 4.1 (Continued)



PLANE ARRIVAL
50
51
52
53
54
55
56
57
58
60
61
62
63
64
65
67
71
72
75
76
77
81
82
83
84
85
86
87
88
89
90
132
133
135
141
147
155
156
157
161
162
163
164
PARK
DOCUMENT TRANSFER
PIVOT NOSE AND K-LOADER
PIVOT NOSE UNLOAD
K-LOADER
K-LOADER AND PIVOT NOSE
CRANE AND PIVOT NOSE
CHECK INSPECT CODE
CRANE AND K-LOADER
25 LINES
9 LINES
TRANSFER TO:
PALLET STORAGE
MULTIPALLET STORE
OVERSIZE STORE
PLANE DOCK
59
10
29
35
PRIORITY
TRANSFER TO: TRUCK DOCK
TRFR TO PALLET BREAKDOWN
PALLET BREAKDOWN
SORT
CHECK
TRFR TO RECRATE AREA
RECRATE
TRANSFER TO:
STACKER 20
TRUCK 81
BUILDUP 23
TRANSFER TO:
TRUCK 81
STACKER 20
BUILDUP 23
CARTABLE LOAD
MULTIPALLET LOAD
TRUCK DEPARTURE
PALLET OVERSIZE LOAD
∞

Figure 4.1 (Continued)

ivities associated with all of the branches are given on the diagram.

Some of the assumptions made to simplify the model are: (1) Only freight which can be palletized is handled inside the terminal. (2) Trucks carry 15 ton loads but are unloaded or loaded in 3 ton groups. (3) Airplanes arrive and depart with 75 tons for a C-5A and 27 tons for a C-141. (4) Three tons of unloaded truck cargo is assumed to be of the same type and thus is processed through the model, intact, as a multipallet, pallet, etc. (5) One pallet of unloaded plane cargo is assumed to be of the same type and is processed as an entity. (6) There are three accumulation nodes (36, 38, 81) used to represent cargo buildup until a full load for a plane or truck is stored before the loading operation is initiated. (7) Arrivals are generated by a time distribution. (8) The terminal operation is based upon a 24 hour work day, but can be extended.

Bottlenecks of the air cargo terminal operations are shown in Figure 4.1 by means of the isolated nodes: 106, 107, 108, 115, 116, 117, 126, 127, 128, 129, 130, 131, 132, 133, 135, 141, 147, and 149.

The GERTSIIIQ simulation model also has imbedded within it activities which are a direct function of construction specifications. These are: (1) The distance between stations determines the transfer times. Thus either the distance between stations can be reduced, or the speed of transporting cargo increased, (i.e. by employing a faster conveyor or towline.) (2) The waiting area needed to accommodate the service determines the maximum allowable number in a queue node. Since the server can service items only at a specified rate, the number of servers may have to be increased if saturation occurs. Thus additional space and equipment may be needed to accommodate the required arrival and departures to the queue. (3) The height of the building and the storage area permitted determines the travel time distribution of storage equipment, particularly the stacker cranes. The retrieval and storage time on the GERT IIIQ diagram for each type of cargo is directly related to the dimensions of the storage area.

The model is flexible enough to remove many if not all of the restrictions spelled out above by changing parameters or by modifying the network slightly.

Simulation Studies: The Dortech-type facility modeled in Figure 4.1 was simulated on the IBM 360/91 digital computer at UCLA. All parameter values were calculated from construction drawings of the planned air cargo facility at Travis Air Force Base and from a Dortech report on material handling equipment [3]. (The parameter values are not listed here, but can be obtained in reference [4]).

A series of simulation runs were conducted to determine potential bottlenecks and to evaluate the best construction-oriented strategy that could be taken to correct them. First the model was simulated recursively to ascertain average queue values of steady state operations then ten simulated one-day operations were run to establish average bottleneck values and to justify the probabilities associated with activities emanating from probabilistic nodes.(For twenty plane departures/day, ten simulation runs proved to be adequate.)The results of simulation runs are summarized in Table 4.2 which also lists the bottlenecks that occurred, the queue nodes and branches affected, the average number in the queues, and values of service time.

Once the steady state operating conditions were established, three separate strategies to heuristically eliminate the bottlenecks were obtained. These were: strategy 1 - reducing service times on branches emanating from the queue node bottlenecks; strategy 2 - increasing the maximum number allowed in the bottleneck queues; and strategy 3 - increasing the number of servers (queues) at each bottleneck point. Table 4.3 summarizes the new conditions achieved in eliminating the bottlenecks.

Since construction costs can vary greatly depending upon the strategy adopted, inference was based upon practicality and cost. Strategy 1 would require extremely fast equipment and additional manpower; strategy 2 would require that large storage space be added; strategy 3 would require that additional space, people, and machinery be added. Of all three strategies, only strategy 3 seemed practical and less costly.

Some representative times to process cargo are shown in Table 4.4. Note the time needed to completely load an airplane according to the strategies taken, and the time required to unload cargo at the truck dock. The strategy 1 times are unrealistic; strategy 2 just adds a lot of space without improving the processing times; strategy 3 constitutes a practical solution.

Table 4.5 summarizes the effects of each strategy on the critical queue nodes. Again strategy 3 represents the only feasible solution.

SUMMARY AND CONCLUSIONS

The Dortech simulation model proved to be quite effective in locating air cargo bottlenecks related to construction considerations. Realistic countermeasures can be taken to compensate for any cargo delays. The results were satisfactory.

REFERENCES

1. Pritsker, A.A.B., Happ, W.W., "GERT: Part I - Fundamentals", Journal of Industrial Engineering, Vol. XVII, No. 5, 1966, pp. 267-274.

2. Pritsker, A.A.B., Burgess, R.R., "The GERT Simulation Programs: GERTS III, GERTS IIIQ, GERTS IIIC, and GERTS IIIR", NASA/ERC Contract NAS-12-2113, Virginia Polytechnic Institute, Department of Industrial Engineering, 1970.

3. Air Freight Terminal Complex: Architect-Engineer Definitive Report, Prepared by Dortech, Stamford, Conn., Contract No. F11623-68-C-0411, for Scott Air Force Base, 1968

4. Lee, Chao, "Parametric Studies of an Air Cargo Terminal", M.S. Thesis, School of Engineering and Applied Science, UCLA, Fall, 1971.

TABLE 4.2

Bottleneck Nodes	147	133	128	117	115
Queue Nodes Affected	83-88	162,164 75	32	23,91	16,17
Max. No. Allowed	5	9	4	150	6
No. of Servers	6	3	1	2	2
Branches Affected	83-89 85-89 88-89	164-76	32-33	91-24	17-18
Service Time (Hrs)	3.0	.75	3.0	.75	3.0

TABLE 4.3

Bottlenecks Eliminated By		147	133	128	117	115
Strat. 1	Service Time	.80	.10	.80	.10	.80
Strat. 2	Max. No. Allowed	30	20	10	350	10
Strat. 3	No. of Servers	10	4	2	6	3

TABLE 4.4

Event	Node	Orig. Proc. Time	Strat. 1	Strat. 2	Strat. 3
Plane Depart.	45	30.6 (hrs)	14.1	30.6	19.86
Pallet Unload	48	3.19	.75	3.19	3.26
Priority Unl.	33	7.27	.91	7.51	3.59
Noncar. Unl.	27	6.07	1.06	6.07	5.73
Cart. Unload	18	6.13	.91	6.19	3.91
Multipal. Unl.	8	3.12	.83	3.12	3.07

TABLE 4.5 (a)

Ave. No. In Queue

Queue Node	Original Design	Strat. 1	Strat. 2	Strat. 3
83	3.45	.143	8.26	.1
85	1.12	.00	1.14	.15
164	2.7	.00	3.34	.53
32	1.68	.00	2.23	.0
91	142	101	186	111
17	2.16	.08	2.4	.78

TABLE 4.5 (b)

Ave. Busy Time of Queue

Queue Node	Original Design	Strat. 1	Strat. 2	Strat. 3
83	.89	.43	.89	.1
85	.77	.149	.77	.24
164	.75	.00	.75	.33
32	.87	.33	.87	.52
91	1	1	1	1
17	1	.23	.91	.77

ACTIVITY NETWORKS TO MODEL TRANSPORTATION SYSTEMS SUBJECT TO FACILITY CONSTRAINTS

H. A. PORTE* AND W. W. HAPP**
*U. S. Army Corps of Engineers
Champaign, Illinois 61820
**University of Illinois
Urbana, Illinois 61801

ABSTRACT

Cargo flow through an air cargo terminal is modeled as an activity network by utilizing: (a) deterministic and probabilistic decision-making elements as nodes, (b) activities or branches which relate the nodes and whose characteristics determine the magnitude and delay of commodity flow, and (c) a set of statistical monitors to count events and to perform statistical evaluations at strategic points of the network.

AIR CARGO FACILITY/TRANSPORTATION INTERFACE

The design of an air terminal must meet anticipated traffic demands imposed by the introduction of jumbo jets. This paper is concerned with the cargo handling aspect of the traffic flow in an air freight terminal complex, and its relationship to construction/facility parameters.

The cargo flow pattern is dependent upon such facility parameters as loading docks, landing strips, transfer stations, storage areas, and loading ramps. It will also be necessary to extrapolate current trends and include anticipated major technical advances in several interacting areas, for example (a) Traffic flow and volume will not follow necessarily the natural growth pattern but change radically due to introduction and increased usage of jumbo jets. (b) Materials handling technology has changed cargo flow patterns significantly, as illustrated by the employment of standardized, lightweight shipping containers. (c) Forecasts of technological advances and of traffic demands are customarily made for five years, but the investor may prefer a forecast for thirty years to encompass the anticipated life cycle of the facility.

Previous investigations [1-8] relating traffic flow to facility parameters may be grouped into the following broad categories: (a) Operations research studies of harbors, airports, warehouses, and freight terminals; (1) airport and shipping terminals (2) resource allocations (3) technological innovations and forecasts, (b) Analytical studies on properties of transportation networks; (1) commodity flow, (2) flow optimization, (3) network models, (c) Computer-assisted analysis and control of transportation phenomena; (1) user-oriented programs and programming techniques, (2) analogue and digital simulation of traffic problems, (3) modelling and optimization techniques; (d) Queueing techniques.

The purpose of this paper is to establish a computer simulation model that relates facility parameters to traffic flow into and out of an air cargo terminal. Particular attention is given to the truck offload dock and place dock areas and their effect on the terminal capacity. A model is postulated and simulated by GERTS IIIQ, which is described in the next section. The model is then exercised to forecast deficiencies in design specification of air terminals with respect to projected cargo flow.

THE GERTS IIIQ SIMULATION PROGRAM

GERTS IIIQ, an acronym for Graphical Evaluation and Review Techniques Simulation with a queueing capability, is an engineering oriented stochastic event simulation program [9]. Operating and coding instructions are well documented [10], but are briefly surveyed here to make the article self-contained.

The GERTS IIIQ program models systems of interest in terms of an network representation, i.e. branches and nodes. A summary of the various types of node elements is shown in Tables 1-3. A node is deterministic if all output activities are scheduled; only one output activity is permitted with queue nodes. A node is probabilistic if only one output activity is scheduled, the selection of which is determined by the probability assigned to the output branches. Probabilistic outputs are permitted with queue nodes since only one activity is scheduled.

Branches represent activities, and when initiated, the time for completion is specified according to a time distribution. The following time distributions are presently available: constant, normal, uniform, erlang, lognormal, poisson, beta, gamma, and beta fitted to three parameters as in PERT.

A GERTS IIIQ program simulates a given system by scheduling activities starting with the source nodes until a number of sink nodes are realized. The entire simulation procedure is repeated until all of the user-specified simulations are run, and all the desired statistics accumulated. The statistics which can be collected by GERTS IIIQ event nodes are: (1) the probability a node is realized, (2) the mean, standard deviation, minimum and the maximum number of times that a node is realized, (3) the time to realize a node the first time, (4) the time required to traverse between two specified nodes, (5) the time delay from the completion of the first activity to the time the node is realized again. The statistics which can be collected by GERTS IIIQ queue nodes are (1) average number in the queue, (2) average busy time of processor, and (3) average balkers per unit time. First time statistics are automatically collected on all queue and sink nodes, if not specified by the user.

QUALITATIVE DESCRIPTION OF AN AIR FREIGHT TERMINAL

A relatively simple approximation of a conceptual flow pattern in an air freight terminal is illustrated in Figure 1. Incoming surface cargo is offloaded from trucks, processed through the terminal, and eventually loaded onto a plane; thus all incoming truck cargo departs the terminal as export cargo via air. Incoming air cargo is offloaded from planes, processed through the terminal and either loaded onto a truck or onto another plane; thus incoming plane cargo departs the terminal as retrograde cargo via truck or as transshipped cargo via plane.

A network model of an air freight terminal using the GERTS IIIQ notation is shown in Figure 2. The probabilistic and deterministic decision points or nodes are linked by branches. It is instructive to regard each node as an event counter triggered by activities, which may be commodity flow or information flow through the network.

Node 2 is a source node which initiates the first truck arrival at the beginning of the simulation. Successive truck arrivals are then produced at node 3 which has a feedback loop and acts as a truck arrival generator.

The branch between node 39 and node 4 is a service activity representing the time to offload a truck; this time was given a normal distribution in the computer runs. Node 4 is a probabilistic node which determines what fraction of the incoming surface cargo is in the form of loose packages and what fraction is prepalletized. For the purpose of the runs it was assumed that only cargo on pallets would be loaded onto or off planes, therefore, all package cargo follows the path from node 4 to node 8 to node 16 to node 6 whereas prepalletized cargo move directly from node 4 to node 6. Palletized cargo then continues through the terminal to node 15 where a loaded aircraft departs the terminal. Node 17 is another source node which initiates aircraft arrival and node 18 is a plane arrival generator. Node 32 is a sink node; and the simulation terminates after a predetermined number of planes have arrived. The branch between node 40 and node 19 is a service activity representing the time to offload a plane. Node 19 is a probabilistic node which determines what fraction of incoming plane cargo will be transshipped by moving to node 34 or node 10 and finally through to node 15; the remainder of the incoming plane cargo will leave the terminal on trucks by moving to nodes 24 or 21 and finally to truck departure node 30.

SIMULATION STUDIES

Simulation studies were conducted to evaluate the effects of cargo flow through an air terminal subject to the following conditions: Case 1: unrestricted cargo flow (no facility constraints as shown in Figure 2), Case 2: queues at the truck offload dock (Figure 3 depicts the changes to the GERTS IIIQ network diagram required to add queue nodes to the truck offload dock). Both constant and normally distributive truck arrival times were considered for both cases.

The following assumptions were made: (1) Unrestricted manpower and cargo handling quipment, i.e. no delays were caused within the terminal complex itself through a lack of material handling equipment and men, (2) Constant arrival rates of planes, and (3) No queues internal to the terminal.

Normal Operating Conditions: Referring to Figure 4, the steady-state performance, measured in tons/day, is depicted by the variables: T_i, truck cargo in; P_i, plane cargo in; T_0, truck cargo out; and P_0, plane cargo out. P_i, T_i, and x, the fraction of incoming cargo that is transshipped, are known; and P_0 and T_0 can be calculated by the relationships:

$$P_0 = T_i + P_i x$$

$$T_0 = P_i(1-x)$$

Parametric solutions to these equations are illustrated in Figure 5, for $x = 0.73$.

A series of computer runs was made to simulate terminal operations for one day up to and including ten weeks to determine how long it would take to reach normal operating conditions starting with an empty terminal. These results are shown in Figure 6. Although the time required to reach steady-state conditions is dependent on the activity constraints inside the terminal, its value is not important here; only the input and output values of T_i, P_i, T_0, and P_0 are relevant. The quantity $P_i + P_0$ is commonly referred to as the terminal capacity.

Facility Constraints: The effect of facility constraints such as the number of available truck dock areas on outgoing plane cargo P_0 was examined. This was accomplished by: (1) limiting the number of servers at the truck unloading dock in the GERTS

IIIQ network (Figure 2) - represented by service activity between nodes 39 and 4, and (2) limiting the number of servers at the plane loading dock - represented by the branch between nodes 13 and 14. At most, four servers were added to the network (case 2); the effect of adding a large number of queues to the truck dock (case 1) was also calculated theoretically and used as an upper bound. (Note that the effect of an infinite number of queue nodes is the same as having no queue nodes at all.)

The computer runs were started with the terminal initially empty and allowed to operate for a period of two weeks. The critical time parameters were: (1) truck arrival rate of 24 trucks/day (constant and normally distributed), (2) truck offload time - 3 hours normally distributed between 2.5 and 3.5 hours, (3) plane load time - .5 hours (a constant time), (4) plane arrival rate - 3 planes/day (constant), (5) each truck carries 15 tons, and (6) each plane carries 105 tons.

The results for one to four, and a large number of truck offloading docks are presented in Figure 7. It is obvious that as the number of servers drops below three there is a significant decrease in the amount of outgoing plane cargo P_0 that can be processed. Also the time to offload a truck increases inversely. The average number in the queue and the percentage of busy time of the server also increases as the number of servers decreases, as expected. It was found that ten simulation runs were sufficient to guarantee that the probabilities at the output of probabilistic nodes were within 4% of the specified values shown in Figure 2.

The recommended number of servers is three.

There was no significant effect on P_0 when a queue was inserted at the plane loading dock which means that the dominant bottlenecks occurred at the truck dock. In other words the service time to load a plane was small with respect to the arrival rate at node 13. Constant or normally distributed truck arrival rates produced essentially the same results.

SUMMARY AND CONCLUSIONS

The simulation studies focused on one operations point at which a bottleneck could occur. The effect of corrective construction measures can then be predicted to remedy the situation. Similar studies can also be made to locate potential bottlenecks both within and outside the terminal, and collectively treated, through simulation, to improve the overall system performance. Actual on site investigations are needed to determine the adequacy of the simulation model and to improve it when necessary.

REFERENCES

1. Miller, A. J., "Queueing at Single-Berth Shipping Terminal", Journal of the Waterways, Harbors and Coastal Engineering Division, ASCE, Vol. 97, No. WW1 Proc. Paper 7895 February, 1971, pp. 43-56.
2. Braaksma, J. P. and Shortreed, J. H., "Improving Airport Gate Usage with Critical Path", Transportation Engineering Journal, ASCE, Vol. 97, No. TE 2, Proc. Paper 8010, May, 1971, pp. 187-203.
3. Hoppe, C. W., "Using Simulation To Solve Transportation Problems", Management Controls, Dec. 1970, pp. 253-255.

4. Yechiali, U., and Naor, P., "Queueing Problems with Heterogeneous Arrivals and Service", Operations Research, Vol. 19, No. 3, May 1971, pp. 722-734.
5. Szwarc, "The Transportation Problem with Stochastic Demand", Management Service, 11, 33-50, (1964).
6. Midler, J. L., "A Stochastic Multiperiod Multimode Transportation Model", Trans. Sci., #, 8-30, (1969).
7. Young, R. N., Ockert, W. W., and Pixton, C. E., "Evaluating Transportation Systems", Transportation Engineering Journal of ASCE, Vol. 97, No. TEZ, Proc. Paper 8131, May 1971, pp. 347-352.
8. Carroll, J. L., Bronzini, M. S., "Simulation of Waterway Transport Systems", Transporation Journal of ASCE, Vol. 97, No. TE3, Aug., 1971, pp. 527-539.

TABLE 1: Terminal Nodes

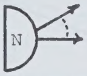
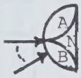
PURPOSE	SYMBOL	PARAMETERS
Source Node		N - Node Number A - NO. OF ACTIVITIES FOR INITIAL REALIZATION B - NO. OF ACTIVITIES FOR SUBSEQUENT REALIZATION
Sink Node		

TABLE 2: Event Nodes

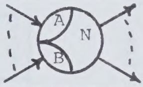
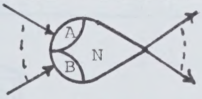
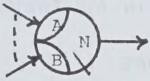
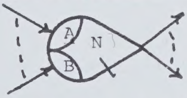
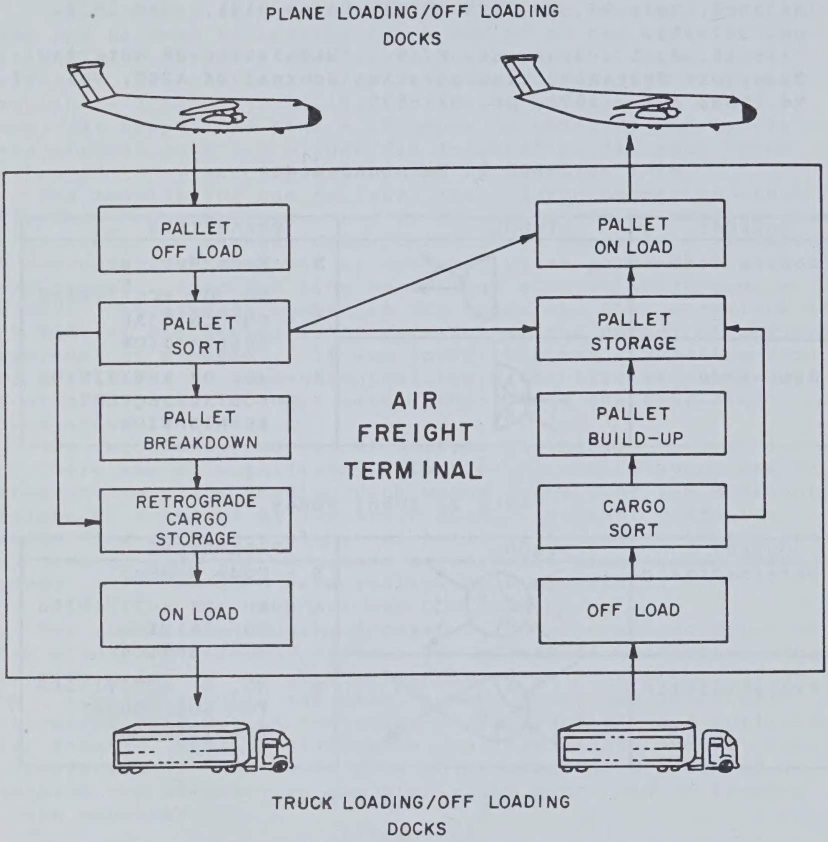
OUTPUT	SYMBOL	PARAMETERS
Deterministic		N - Node Number A - NO. OF ACTIVITIES FOR INITIAL REALIZATION B - NO. OF ACTIVITIES FOR SUBSEQUENT REALIZATION
Probabilistic		

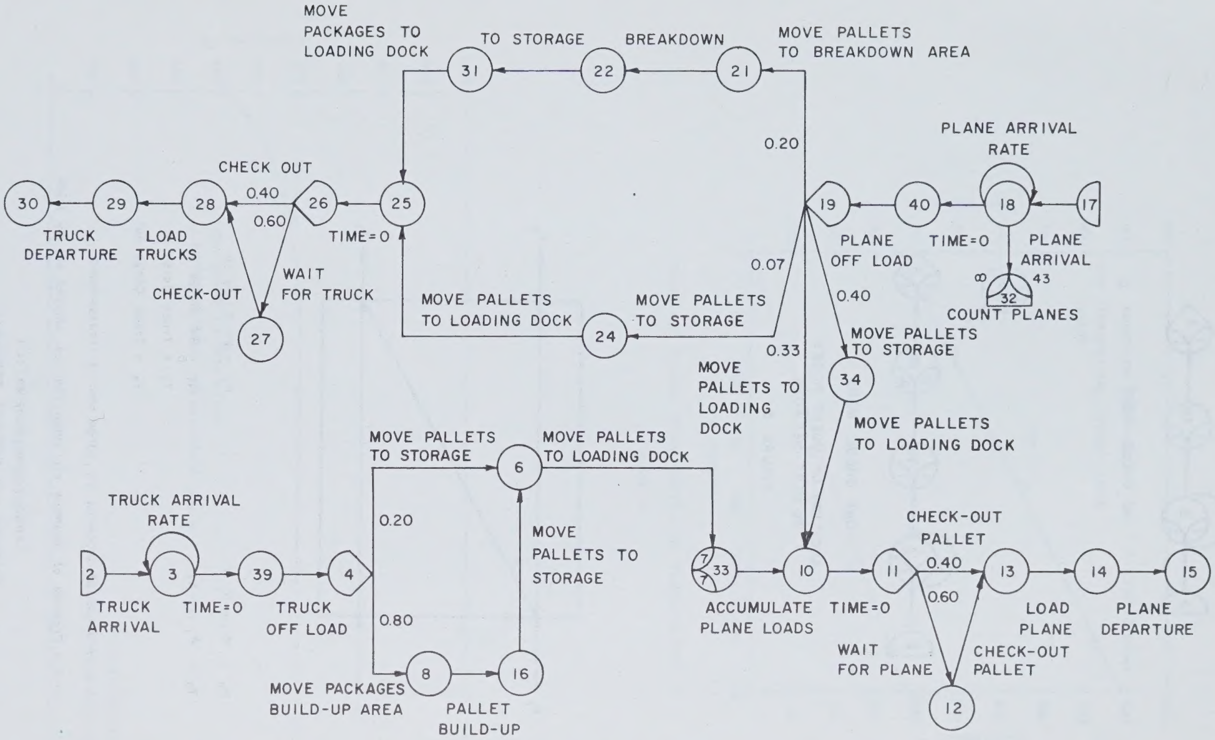
TABLE 3: Queue Nodes

OUTPUT	SYMBOL	PARAMETERS
Deterministic		N - Node Number A - INITIAL NUMBER IN QUEUE B - MAX. NUMBER ALLOWED IN QUEUE
Probabilistic		


PLANE LOADING/OFF LOADING DOCKS
PALLET OFF LOAD
PALLET SORT
PALLET BREAKDOWN
RETROGRADE CARGO STORAGE
ON LOAD
AIR FREIGHT TERMINAL
PALLET ON LOAD
PALLET STORAGE
PALLET BUILD-UP
CARGO SORT
OFF LOAD
TRUCK LOADING/OFF LOADING DOCKS

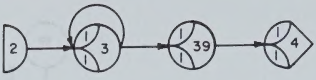
CONCEPTUAL FLOW PATTERN IN AIR FREIGHT TERMINAL

FIGURE I

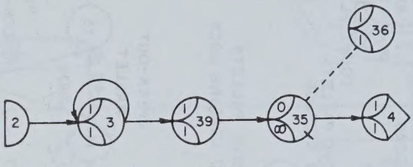

MOVE PACKAGES TO LOADING DOCK
31
TO STORAGE
22
BREAKDOWN
21
MOVE PALLETS TO BREAKDOWN AREA
0.20
PLANE ARRIVAL RATE
17
18
40
19
PLANE OFF LOAD
TIME=0
PLANE ARRIVAL
∞
43
32
COUNT PLANES
0.07
0.40
MOVE PALLETS TO STORAGE
34
MOVE PALLETS TO LOADING DOCK
0.33
MOVE PALLETS TO LOADING DOCK
MOVE PALLETS TO STORAGE
24
MOVE PALLETS TO LOADING DOCK
25
TIME=0
26
CHECK OUT
0.40
0.60
28
29
30
TRUCK DEPARTURE
LOAD TRUCKS
CHECK-OUT
WAIT FOR TRUCK
27
TRUCK ARRIVAL RATE
2
3
39
4
TRUCK ARRIVAL
TIME=0
TRUCK OFF LOAD
0.20
0.80
MOVE PALLETS TO STORAGE
6
MOVE PALLETS TO LOADING DOCK
MOVE PALLETS TO STORAGE
8
16
MOVE PACKAGES BUILD-UP AREA
PALLET BUILD-UP
7
7
33
ACCUMULATE PLANE LOADS
10
TIME=0
11
CHECK-OUT PALLET
0.40
0.60
13
14
15
LOAD PLANE
PLANE DEPARTURE
WAIT FOR PLANE
CHECK-OUT PALLET
12

GERT NETWORK FOR AIR FREIGHT TERMINAL

FIGURE 2


2
1
3
1
1
39
1
1
4
1

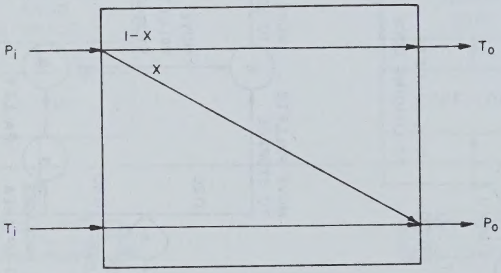
NO QUEUE NODE


2
1
3
1
1
39
1
0
35
∞
1
36
1
1
4
1

ONE QUEUE NODE

ADDITION OF QUEUE NODES
TO GERT NETWORK

FIGURE 3


P_i
$1-X$
X
T_o
T_i
P_o

$P_o = T_i + P_i X$

$T_o = P_i (1-X)$

P_i = Air Cargo in

P_o = Air Cargo out

T_i = Truck Cargo in

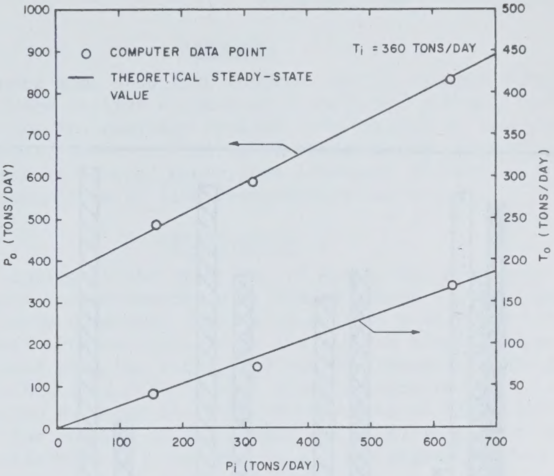
T_o = Truck Cargo out

X = Fraction of incoming air cargo that is transshipped

$1-X$ = Fraction of incoming air cargo that is shipped out by truck

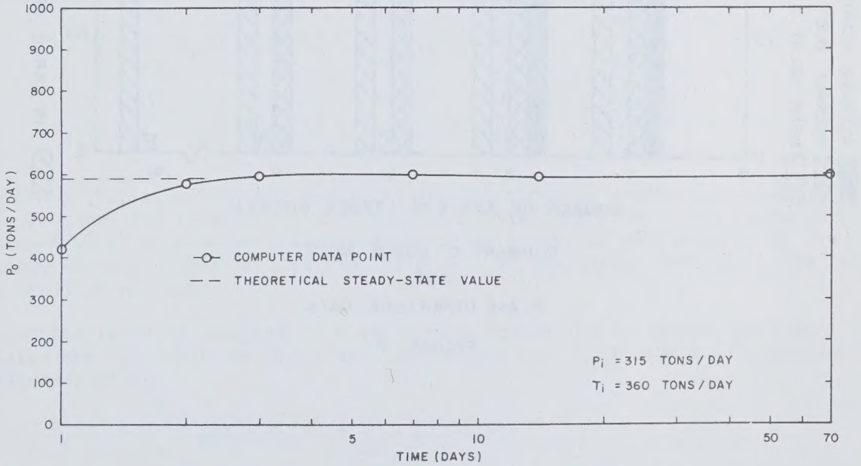
CARGO FLOW (SCHEMATIC)
THROUGH AIR FREIGHT TERMINAL

FIGURE 4



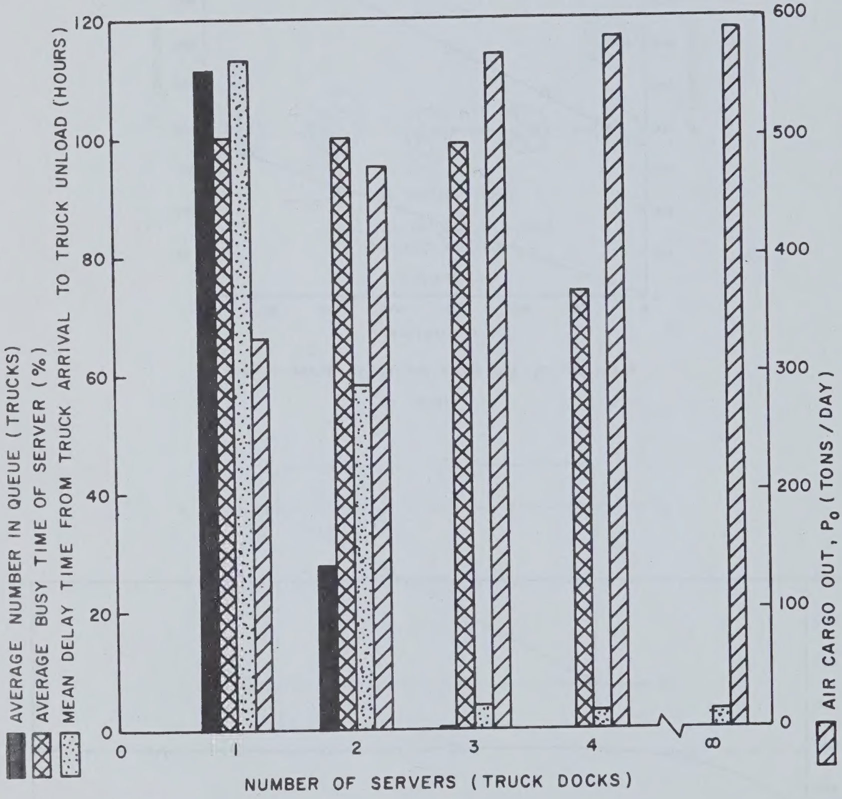
PLANE OUT (P_o) AND TRUCK OUT (T_o) VS. PLANE IN (P_i)

FIGURE 5



PLANE OUT (P_o) VS TIME

FIGURE 6



SUMMARY OF QUEUE NODE
AND
PLANE DEPARTURE DATA

FIGURE 7

STABILITY OF DISTRIBUTED NETWORKS WITH NONLINEAR ELEMENTS

PAUL J. SCHWYN
I.B.M. Corporation
Endicott, New York 13760

THEODORE A. BICKART
Syracuse University
Dept. of Electrical & Computer Eng.
Syracuse, New York 13210

ABSTRACT

This paper presents some stability theorems for an integro-differential equation which characterizes a class of distributed networks with nonlinear elements. Then the theorems are applied to a particular category of such electrical networks, those consisting of transmission lines interconnected with lumped linear and lumped memoryless nonlinear elements. A simple example with a nonuniform RC line demonstrates the results.

INTRODUCTION

This paper considers the stability of electrical networks consisting of transmission lines interconnected with lumped linear elements and lumped memoryless nonlinear elements. The value of such a stability theory derives from the fact that networks containing distributed elements have gained increasing importance with the widespread use of integrated circuits. Furthermore, interconnections of today's high speed integrated circuits can no longer be neglected as short circuits, but have to be considered as transmission lines. The results of this paper are an extension of those obtained by Willson [1] and Prada [2] and are to a large degree derived from [3].

FUNCTIONAL REPRESENTATION OF TRANSMISSION LINES

The distributed elements considered in this paper are transmission lines. The lines are embedded in some simple resistive two-ports as shown in Figure 1. A two-port of class a has the voltages $v_o(t)$ and $v_\ell(t)$ as inputs; the outputs are the currents $i_o(t)$ and $i_\ell(t)$. Let v_{ot} and $v_{\ell t}$ denote the translations of $v_o(\tau)$ and $v_\ell(\tau)$ defined on $(-\infty, t]$ to the interval $(-\infty, 0]$. Furthermore, let v_{ot} and $v_{\ell t}$ be continuous and bounded on $(-\infty, 0]$. Then the currents $i_o(t)$ and $i_\ell(t)$ are characterized by the following functional equations:

$$i_o(t) = \int_{-\infty}^{0} v_{ot}(\tau)d_\tau g_\sigma(\tau) + \int_{-\infty}^{0} v_{\ell t}(\tau)d_\tau g_\delta(\tau)$$

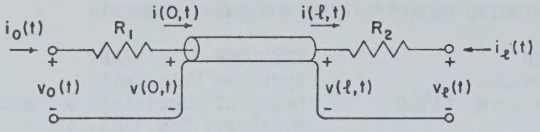
$$i_\ell(t) = \int_{-\infty}^{0} v_{ot}(\tau)d_\tau g_\mu(\tau) + \int_{-\infty}^{0} v_{\ell t}(\tau)d_\tau g_\epsilon(\tau), \tag{1}$$

where $g_\sigma(t)$, $g_\delta(t)$, $g_\mu(t)$, and $g_\epsilon(t)$ are functions of bounded variation on $(-\infty, 0]$ and the integral is of the Riemann-Stieltjes type. This representation exists for a large class of distributed lines, in particular, for the lossy distortionless LC line and, if R_1 and R_2 are larger than zero, for the nonuniform RC line.

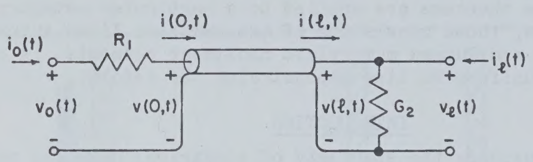
Similarly, a two-port of class b (See Figure 1b.), for which $v_o(t)$ and $i_\ell(t)$ are the input variables and the outputs are $i_o(t)$ and $v_\ell(t)$, has the representation:

$$i_o(t) = \int_{-\infty}^{0} v_{ot}(\tau)d_\tau g_\sigma(t) - \int_{-\infty}^{0} i_{\ell t}(\tau)d_\tau g_\delta(\tau)$$

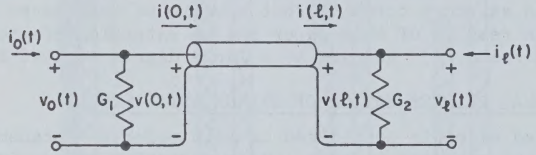
$$v_\ell(t) = - \int_{-\infty}^{0} v_{ot}(\tau)d_\tau g_\mu(\tau) + \int_{-\infty}^{0} i_{\ell t}(\tau)d_\tau g_\epsilon(\tau). \tag{2}$$


$i_0(t)$
R_1
$i(0,t)$
$i(\ell,t)$
R_2
$i_\ell(t)$
$v_0(t)$
$v(0,t)$
$v(\ell,t)$
$v_\ell(t)$
+
-

(a) CLASS a


$i_0(t)$
R_1
$i(0,t)$
$i(\ell,t)$
$i_\ell(t)$
$v_0(t)$
$v(0,t)$
$v(\ell,t)$
G_2
$v_\ell(t)$
+
-

(b) CLASS b


$i_0(t)$
$i(0,t)$
$i(\ell,t)$
$i_\ell(t)$
$v_0(t)$
G_1
$v(0,t)$
$v(\ell,t)$
G_2
$v_\ell(t)$
+
-

(c) CLASS c

FIG. 1 TRANSMISSION LINE TWO-PORTS

If the transmission line is an RC line, then R_1 is required to be larger than zero for the above representation to exist.

Finally, a two-port of class c (See Figure 1c.), for which $i_o(t)$ and $i_\ell(t)$ are the input variables and the outputs are $v_o(t)$ and $v_\ell(t)$, has the representation:

$$v_o(t) = \int_{-\infty}^{0} i_{ot}(\tau) d_\tau g_\sigma(\tau) + \int_{-\infty}^{0} i_{\ell t} d_\tau g_\delta(\tau)$$

$$v_\ell(t) = \int_{-\infty}^{0} i_{ot}(\tau) d_\tau g_\mu(\tau) + \int_{-\infty}^{0} i_{\ell t} d_\tau g_\epsilon(\tau). \tag{3}$$

The g-functions for the above equations are closely related to the step response of the two-ports. For the stability criteria established, only the discontinuity $g(0) - g(0-)$ of g at $t = 0$ and the total variation $\vartheta(g(0-))$ of $g(t)$ over the interval $(-\infty, 0)$ are of interest. (Note: The interval is open and excludes the discontinuity at zero.) As an example the relevant values for the g-functions of a two-port of class a with a non-uniform RC line are listed. In the following Table 1, $R_\Sigma = R_1 + R_2 + R_o$, where R_o is the total line resistance.

		$g_\eta(0)-g_\eta(0-)$	$\nu(g_\eta(0-))$	$g_\eta(0)$
η	σ	$\frac{1}{R_1}$	$\frac{1}{R_1} - \frac{1}{R_\Sigma}$	$\frac{1}{R_\Sigma}$
	δ	0	$\frac{1}{R_\Sigma}$	$\frac{1}{R_\Sigma}$
	μ	0	$\frac{1}{R_\Sigma}$	$\frac{1}{R_\Sigma}$
	ϵ	$\frac{1}{R_2}$	$\frac{1}{R_2} - \frac{1}{R_\Sigma}$	$\frac{1}{R_\Sigma}$

Table 1 g-functions of nonuniform RC lines of class a.

A CLASS OF NETWORKS WITH DISTRIBUTED AND NONLINEAR ELEMENTS

The networks for which stability theorems will be developed can be represented as in Figure 2. They consist of the following time invariant

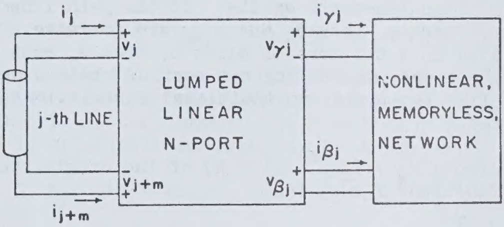


FIG. 2 NETWORK BLOCK DIAGRAM

components: (1) lumped, linear elements; (2) lumped, memoryless, nonlinear elements; and (3) distributed parameter elements — a collection of transmission lines.

The nonlinear multiport in Figure 2 has p-ports where the currents are the independent variables and q ports where the voltages are the independent variables. Let $\underline{i}_\gamma$ and $\underline{v}_\beta$ be p- and q-vectors whose components are these independent variables and let $\underline{v}_\gamma$ and $\underline{i}_\beta$ be p- and q-vectors of the complimentary dependent variables, specified by the nonlinear vector valued function:

$$\begin{bmatrix} \underline{v}_\gamma \\ \underline{i}_\beta \end{bmatrix} = \underline{\mathcal{F}} \begin{bmatrix} \underline{i}_\gamma \\ \underline{v}_\beta \end{bmatrix} \tag{4}$$

The linear lumped part of the network has p + q + 2m ports, m being the number of distributed lines, and is described by the state equations

$$\dot{\underline{x}}(t) = A\underline{x}(t) + B\underline{u}(t)$$

$$\underline{w}(t) = C\underline{x}(t) + D\underline{u}(t), \qquad (5)$$

where $\underline{x}(t)$ is an n-vector whose elements are the state variables of the network, where the components of the p+q+2m vector $\underline{u}$ are the components of $\underline{v}_\gamma$ and $\underline{i}_\beta$ plus the 2m output variables of the transmission line two-ports, and where $\underline{w}$ is a p+q+2m vector whose components are those of $\underline{i}_\gamma$ and $\underline{v}_\beta$ plus the 2m input variables of the transmission line two-ports.

The following assumptions are made with respect to the system described by (4) and (5):

If the j-th line is embedded in a two-port of class a, then u_{p+q+j} and $u_{p+q+j+m}$ are currents, and w_{p+q+j} and $w_{p+q+j+m}$ are voltages. If the j-th line belongs to a two-port of class b, then u_{p+q+j} and $w_{p+q+j+m}$ are currents, and $u_{p+q+j+m}$ and w_{p+q+j} are voltages. Finally, if the j-th line is in a two-port of class c, then u_{p+q+j} and $u_{p+q+j+m}$ are voltages, and w_{p+q+j} and $w_{p+q+j+m}$ are currents. (6)

D is a (p+q+2m) x (p+q+2m) diagonal matrix with nonnegative elements $d_{\gamma 1}, \ldots, d_{\gamma p}, d_{\beta 1}, \ldots, d_{\beta q}, d_1, \ldots, d_{2m}$. Furthermore, the elements $d_1 \ldots d_{2m}$ are such that the functional representation for each of the transmission line two-ports exists: If the j-th line is an RC line in a two-port of class a, then d_j and d_{j+m} are positive. If the j-th line is an RC line in a two-port of class b, then d_j must be positive. For the remaining cases, an RC line two-port of class c and lossy distortionless LC line two-ports, no additional restrictions on d_j and d_{j+m} are required. (7)

In some neighborhood $N_H = \{\underline{x} \epsilon E^n : \|\underline{x}\| < H\}$ of the origin, there exists a vector valued function $\underline{F}$ such that

$$\begin{bmatrix} \underline{v}_\gamma \\ \underline{i}_\beta \end{bmatrix} = \underline{F}\,(C_I \underline{x}),$$

where C_I is a submatrix of C consisting of the first p+q rows of the matrix C. The function $\underline{F}$ satisfies a Lipschitz condition on the neighborhood of definition. (8)

For what follows it is convenient to partition the matrices B and C according to the different classes of ports. Let

$$B = [B_I \ \ B_{II} \ \ B_{III}]$$

$$B_I = [B_\gamma \ \ B_\beta], \quad B_{II} = [\underline{b}_1 \ \ \underline{b}_2 \ \cdots \ \underline{b}_m], \quad B_{III} = [\underline{b}_{m+1} \ \cdots \ \underline{b}_{2m}]$$

and

$$C^T = [C_I \ \ C_{II} \ \ C_{III}]$$

$$C_I^T = [C_\gamma^T \ \ C_\beta^T], \quad C_{II}^T = [\underline{c}_1 \ \ \underline{c}_2 \ \cdots \ \underline{c}_m], \quad C_{III}^T = [\underline{c}_{m+1} \ \cdots \ \underline{c}_{2m}],$$

where B_γ and B_β are nxp and nxq submatrices, where C_γ and C_β are pxn and qxn submatrices, and where $\underline{b}_j$ and $\underline{c}_j$ are n-vectors for j = 1, 2,...,2m.

AN INTEGRO-DIFFERENTIAL EQUATION

Combining the functional representation of the transmission line two-ports with the state equations (5) results in the following integro-differential equation describing the network represented in Figure 1:

$$\begin{aligned}\dot{\underline{x}}(t) &= \underline{f}(\underline{x}_t)\\ &= A\underline{x}(t) + B_I\underline{F}(C_I\underline{x}(t)) - \sum_{j=1}^{m}\ \sum_{\eta\in\Delta_j} \underline{b}_{j+\xi_\eta}\,\underline{c}^T_{j+\nu_\eta} \int_{-\infty}^{0} \underline{x}_t(\tau)[d_\tau g_{\eta j}(\tau)],\end{aligned} \tag{9}$$

where $\underline{x}_t$ is the translation of $\underline{x}$ as defined on $(-\infty, t]$ to the interval $(-\infty,0]$ and where the integral is again of the Riemann-Stieltjes type. Note: The integral will exist if $\underline{x}_t$ is a continuous and bounded function mapping $(-\infty,0]$ into E^n. ξ and ν are functions defined on the set $\Delta = \{\sigma, \delta, \mu, \epsilon\}$ as

$$\begin{aligned}\nu_\eta &= 0, \ \eta \in \{\sigma,\mu\} &\quad\text{and}\quad \xi_\eta &= 0, \ \eta \in \{\sigma,\delta\}\\ &= m, \ \eta \in \{\delta,\epsilon\} & &= m, \ \eta \in \{\mu,\epsilon\}.\end{aligned}$$

The set Δ_j is defined as $\Delta_j = \{\eta \in \Delta : \underline{b}_{j+\xi_\eta}\,\underline{c}^T_{j+\nu_\eta} \neq 0\}$.

Let $\nu(g_{\eta j}(t))$ denote the total variation of $g_{\eta j}(t)$ defined on $(-\infty, 0]$ with $\nu(g_{\eta j}(-\infty)) = 0$. Then there exists a function $h(t)$, which is continuous and nondecreasing on the interval $(-\infty, 0]$, such that

$$h(-\infty) = 0$$

and

$$h(t) > \nu(g_{\eta j}(t))$$

for $t \leq 0$ and for every $\eta \in \Delta$ and $j = 1, 2, \ldots, m$.

The nonlinear function $\underline{f}(\underline{x}_t)$ in (9) is defined on $C((-\infty, 0], E^n)$, the space of bounded, continuous functions mapping $(-\infty, 0]$ into E^n. The topology for C is taken to be the compact open topology [4]. It can be shown [1] that this space is metrizable and complete with the metric

$$\rho(\varphi,\psi) = \sum_{k=0}^{\infty} m_k, \quad m_k = \min\{b^k, \sup\|\varphi(t) - \psi(t)\| : t_{k+1} \leq t \leq t_k\},$$

where $\{t_k\}$ is a sequence of real numbers such that $0 = t_0 > \ldots > t_k > t_{k+1} > \ldots$ and such that $\lim_{k\to\infty} t_k = -\infty$ and where $0 < b < 1$. Choosing the sequence $\{t_k\}$ such that $h(t_k) < b^{kk}$ for $k > 0$, then it can be shown that $\underline{f}(\underline{x}_t)$ is Lipschitz continuous on C_H, where $C_H = \{\underline{\varphi} \in C : \|\underline{\varphi}(t)\| < H,\ t \in (-\infty, 0]\}$. Thus, (9) has a unique solution on C_H.

STABILITY THEOREMS

The solution $\underline{x} = \underline{0}$ of equation (9) is said to be stable if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\underline{\varphi} \in C_H$ and $\rho(\underline{\varphi}, \underline{0}) < \delta$ imply that for all $t \geq \sigma$ the solution $\underline{x}_t(\sigma, \underline{\varphi})$ exists and $\rho(\underline{x}_t(\sigma, \underline{\varphi}), \underline{0}) < \epsilon$. If in addition $\rho(\underline{x}_t(\sigma,\varphi), \underline{0}) \to \underline{0}$ as $t \to \infty$, then the solution is said to be asymptotically stable. The solution is said to be unstable if it is not stable. Note: $\underline{x}(\sigma, \underline{\varphi})$ denotes the solution of (9) with initial function $\underline{\varphi}$ at time $t = \sigma$ and $\underline{x}_t(\sigma,\underline{\varphi})$ is the translation to the interval $(-\infty, 0]$ of $\underline{x}(\sigma,\underline{\varphi})$ restricted to $(-\infty, t]$.

The following theorems were developed by using Liapunov theory as extended to functional differential equations (See [5] and [6].). Let P be a positive definite, symmetric $n \times n$ matrix, let the $\beta_{\eta j}$ be positive real numbers, and let $\underline{\varphi} \in C_H$, then the functional

$$V(\underline{\varphi}) = \underline{\varphi}^T(0)P\underline{\varphi}(0) + \sum_{j=1}^{m} \sum_{\eta\in\Delta j} \beta_{\eta j}\int_{-\infty}^{0} \underline{\varphi}^T(\tau)P\underline{c}_{j+\nu_\eta}\underline{c}^T_{j+\nu_\eta} P\underline{\varphi}(\tau)[d_\tau\upsilon(g_{\eta j}(\tau))] \qquad (10)$$

is continuous on C_H. Using this functional and applying the results obtained by Hale [6], the following theorem can be proved:

Theorem 1.

If there exist real numbers $\beta_{\eta j} > 0$, $j = 1, 2, \ldots, m$ and $\eta \in \Delta_j$, and a symmetric, positive definite nxn matrix P such that

$$\underline{x}^T[2PA - \sum_{j=1}^{m} \sum_{\eta\in\Delta j} \{2(g_{\eta j}(0) - g_{\eta j}(0-))P\underline{b}_{j+\xi_\eta}\underline{c}^T_{j+\nu_\eta}$$

$$- (\beta_{\eta j}P\underline{b}_{j+\xi_\eta}\underline{b}^T_{j+\xi_\eta} P + \frac{1}{\beta_{\eta j}} \underline{c}_{j+\nu_\eta}\underline{c}^T_{j+\nu_\eta})\upsilon(g_{\eta j}(0-))\}]\underline{x} + 2\underline{x}^TPB_I\underline{F}(C_I\underline{x}) \leq 0 \qquad (11)$$

in some neighborhood of the origin, then the solution $\underline{x} = \underline{0}$ of (9) is stable. Moreover, if the equality occurs only when $\underline{x} = \underline{0}$, then the solution is asymptotically stable.

This theorem is general and can be applied to all transmission line two-ports having step responses of bounded variation, in particular, two-ports with lossy distortionless LC lines [3] and nonuniform RC lines. Restricting the distributed elements to RC lines only, another stability criterion can be proved.

The matrix $(\underline{b}\,\underline{c}^T)$ can be written as the sum $(\underline{b}\bar{c})_s + (\underline{b}\,\underline{c}^T)_p - (\underline{b}\,\underline{c}^T)_n$, where

$$(\underline{b}\,\underline{c}^T)_s = \frac{1}{2}\underline{b}\,\underline{c}^T - \frac{1}{2}\underline{c}\,\underline{b}^T$$

$$(\underline{b}\,\underline{c}^T)_p = \frac{1}{4}\{\underline{b}\,\underline{b}^T + \underline{b}\,\underline{c}^T + \underline{c}\,\underline{b}^T + \underline{c}\,\underline{c}^T\}$$

$$(\underline{b}\,\underline{c}^T)_n = \frac{1}{4}\{\underline{b}\,\underline{b}^T - \underline{b}\,\underline{c}^T - \underline{c}\,\underline{b}^T + \underline{c}\,\underline{c}^T\}.$$

$(b\ c^T)_p$ and $(\underline{b}\ \underline{c}^T)_n$ are positive semidefinite symmetric matrices and $(b\ c^T)_s$ is a skew symmetric matrix which can be written as

$$(\underline{b}\,\underline{c}^T)_s = W^T \underset{\longleftarrow\; n \;\longrightarrow}{\left[\begin{array}{cc:c} 0 & \mu & \\ -\mu & 0 & 0 \\ \hdashline \multicolumn{2}{c:}{0} & 0 \end{array}\right]} W,$$

where W is a real nxn matrix such that $WW^T = \underline{I}$ and where μ is a positive, real number. Next, the positive semidefinite matrix E is defined as

$$E = W^T \underset{\longleftarrow\; n \;\longrightarrow}{\left[\begin{array}{cc:c} \mu & 0 & 0 \\ 0 & \mu & \\ \hdashline \multicolumn{2}{c:}{0} & 0 \end{array}\right]} W.$$

In the following theorem the pairs $E_{\sigma j}$ and $(P\underline{b}_j\underline{c}^T_j)_s$, $E_{\delta j}$ and $(P\underline{b}_j\underline{c}^T_{j+m})_s$, $E_{\mu j}$ and $(P\underline{b}_{j+m}\underline{c}^T_j)_s$, and $E_{\epsilon j}$ and $(P\underline{b}_{j+m}\underline{c}^T_{j+m})_s$ are defined in the same way as the pair E and $(\underline{b}\ \underline{c}^T)$ above.

Using a Liapunov functional similar to (10), a stability theorem applying only to nonuniform RC lines in two-ports of class a can be proved. In this theorem the resistance $R_{\Sigma j} = d_j + d_{j+m} + R_{oj}$, where d_j and d_{j+m} are the terminating resistances in the two-port and R_{oj} is the total line resistance.

Theorem 2.

If there exists a positive definite symmetric nxn matrix P such that

$$\underline{x}^T[PA - \sum_{j=1}^{m} (\frac{1}{d_j} P\underline{b}_j\underline{c}_j^T + \frac{1}{d_{j+m}} P\underline{b}_{j+m}\underline{c}_{j+m}^T) + \sum_{j=1}^{m} \{(\frac{1}{d_j} - \frac{1}{R_{\Sigma j}})(E_{\sigma j} + (P\underline{b}_j\underline{c}_j^T)_p) +$$

$$+ (\frac{1}{d_{j+m}} - \frac{1}{R_{\Sigma j}})(E_{\epsilon j} + (P\underline{b}_{j+m}\underline{c}_{j+m}^T)_p) + \frac{1}{R_{\Sigma j}}(E_{\delta j} + E_{\mu j} + (P\underline{b}_j\underline{c}_{j+m}^T)_p + (P\underline{b}_{j+m}\underline{c}_j^T)_p)\}]\underline{x}$$

$$+ \underline{x}^T PB_I\underline{F}(C_I\underline{x}) \leq 0 \tag{12}$$

for all $\underline{x}$ in a neighborhood of the origin, then the network in Figure 2 with all the lines being nonuniform RC lines embedded in two-ports of class a is stable. If the equality occurs only when $\underline{x} = \underline{0}$, then the network is asymptotically stable.

In a similar manner two instability theorems can be developed.

Theorem 3

If there exist a positive definite, symmetric nxn matrix P and real numbers $\beta_{\eta j} > 0$, for $j = 1, 2, \ldots, m$ and $n \in \Delta j$, such that

$$\underline{x}^T[2PA - \sum_{j=1}^{m} \sum_{\eta\in\Delta j} \{2(g_{\eta j}(0) - g_{\eta j}(0-))P\underline{b}_{j+\xi_\eta}\underline{c}_{j+\nu_\eta}^T$$

$$+ (\beta_{\eta j} P\underline{b}_{j+\xi_\eta}\underline{b}_{j+\xi_\eta}^T P + \frac{1}{\beta_{\eta j}} \underline{c}_{j+\nu_\eta}\underline{c}_{j+\nu_\eta}^T)\upsilon(g_{\eta j}(0-))\}]\underline{x} + 2\underline{x}^T PB_I\underline{F}(C_I\underline{x}) > 0 \tag{13}$$

for all $\underline{x}$ in a neighborhood of the origin and $\underline{x} \neq \underline{0}$, then (9) is unstable.

Theorem 4

If there exists a positive definite, symmetric nxn matrix P such that

$$\underline{x}^T[PA - \sum_{j=1}^{m} (\frac{1}{d_j} P\underline{b}_j\underline{c}_j^T + \frac{1}{d_{j+m}} P\underline{b}_{j+m}\underline{c}_{j+m}^T)$$

$$- \sum_{j=1}^{m} \{(\frac{1}{d_j} - \frac{1}{R_{\Sigma j}})(E_{\sigma j} + (P\underline{b}_j\underline{c}_j^T)_n) + (\frac{1}{d_{j+m}} - \frac{1}{R_{\Sigma j}})(E_{\epsilon j} + (P\underline{b}_{j+m}\underline{c}_{j+m}^T)_n)$$

$$+ \frac{1}{R_{\Sigma j}} (E_{\delta j} + E_{\mu j} + (P\underline{b}_j\underline{c}_{j+m}^T)_n) + (P\underline{b}_{j+m}\underline{c}_j^T)_n)\}]\underline{x} + \underline{x}^T PB_I\underline{F}(C_I\underline{x}) > 0 \tag{14}$$

for all $\underline{x}$ in a neighborhood of the origin and $\underline{x} \neq \underline{0}$, then the network in Figure 2 with all the lines being nonuniform RC lines embedded in two-ports of class a is unstable.

EXAMPLE

The example considered is the network in Figure 3. The nonlinear element is characterized by the equation $i = f(v)$, where we assume that f satisfies a Lipschitz condition in some neighborhood of the origin, and that $f(0) = 0$. For this network, the state equations (5) are

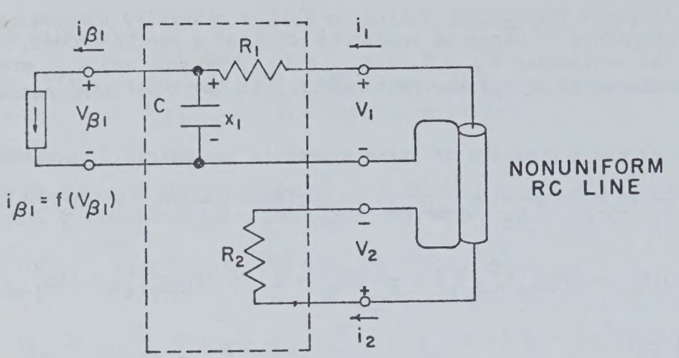


FIG. 3 NETWORK FOR EXAMPLE

$$\dot{x}_1(t) = [0]x_1(t) + [-\tfrac{1}{C}\ \tfrac{1}{C}\ 0] \begin{bmatrix} i\beta_1 \\ i_1 \\ i_2 \end{bmatrix}$$

$$\begin{bmatrix} v_{\beta_1} \\ v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} x_1(t) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & R_1 & 0 \\ 0 & 0 & R_2 \end{bmatrix} \begin{bmatrix} i_{\beta_1} \\ i_1 \\ i_2 \end{bmatrix} .$$

Thus, $A = [0]$, $\underline{b}_{\beta_1} = [-1/C]$, $\underline{b}_1 = [1/C]$, $\underline{b}_2 = [0]$, $\underline{c}_{\beta_1} = [1]$, $\underline{c}_1 = [1]$, $\underline{c}_2 = [0]$, $D_I = [0]$, $d_1 = R_1$, $d_2 = R_2$.

Let P be a positive number and set $\beta_\eta = C/P$ for $\eta \in \Delta_j$ in Theorems 1 and 3. Then, upon using the g-function variations in Table 1 the following stability criteria evolve:

From both Theorem 1 and from Theorem 2, the network is stable if

$$-x_1^2 \frac{1}{R_\Sigma} - x_1 f(x_1) \leq 0 \tag{15}$$

in a neighborhood of the origin.

From Theorem 3, the network is unstable if

$$- x_1^2 \left(\frac{2}{R_1} - \frac{1}{R_\Sigma} \right) - x_1 f(x_1) > 0 \tag{16}$$

for all x_1 in a neighborhood of zero and $x_1 \neq 0$.

From Theorem 4, the network is unstable if

$$-x_1^2 \frac{1}{R_1} - x_1 f(x_1) > 0 \tag{17}$$

for all x_1 in a neighborhood of zero and $x_1 \neq 0$.

In Figure 4 the different sectors of stability are shown. Theorem 4 gives

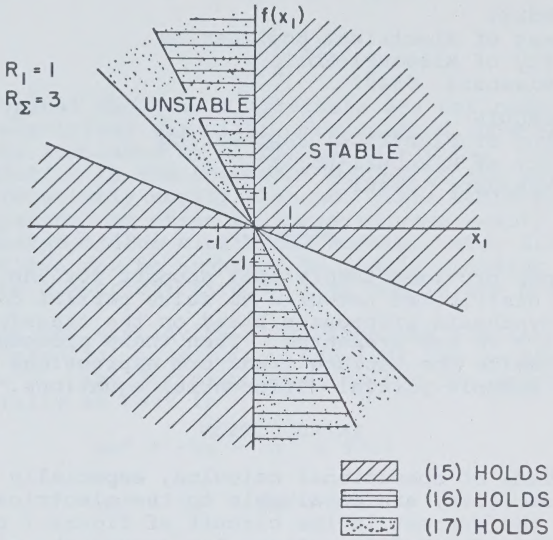


FIGURE 4 STABILITY OF THE NETWORK
IN FIGURE 3

a stronger condition than Theorem 3, which isn't surprising since Theorem 4 applies only to RC lines wereas Theorem 3 applies to a larger class of lines.

For examples of networks with different types of lines consult [3].

REFERENCES

[1] Willson, A.N., A Stability Theory for Nonlinear Distributed Networks, Doctoral Dissertation, Syracuse University, Syracuse, N.Y., May 1967.

[2] Prada, G. and Bickart, T.A., "Stability of Electrical Networks Containing Distributed RC Components," J. Math. Anal. and Appl., vol. 33, 1971, pp. 367-401.

[3] Schwyn, P.J. Stability of Nonlinear Distributed Networks, Doctoral Dissertation, Syracuse University, Syracuse, N.Y., June 1971.

[4] Arens, R.F., "A Topology for Spaces of Transformations," Ann. of Math., vol. 47, 1946, pp. 480-495.

[5] LaSalle, J. and Lefschetz, S., Stability by Liapunov's Direct Method with Applications, Academic Press, New York, 1961.

[6] Hale, J., "Sufficient Conditions for Stability and Instability of Autonomous Functional Differential Equations," Journal of Diff. Equ., vol. 1, 1965, pp. 452-482.

APPLICATION OF DISTRIBUTED NETWORK SYNTHESIS PROCEDURES TO THE SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

R.D. HENSLEY
Department of Electrical Engineering
University of Missouri-Rolla
Rolla, Missouri 65401

J.J. BOURQUIN
Department of Electrical Engineering
University of Missouri-Rolla
Rolla, Missouri 65401

ABSTRACT

This paper provides theoretical support for the idea of constructing distributed networks to solve partial differential equations. Synthesis procedures based on the transformations $W = e^{\sqrt{s}}$ and $\theta = \sqrt{s}$ are presented. Then these procedures are used to synthesize the Laplace transform expressions of the solutions to example partial differential equations.

INTRODUCTION

The methods of operational calculus, especially Laplace transform techniques, are invaluable to the electrical engineer. For example, the current in the circuit of figure 1 can be found by writing Kirchoff's voltage law around the single loop of the circuit and solving the resulting integro-differential equation.

$$L \frac{di}{dt} + Ri + \frac{1}{C} \int_0^t i \, dt = v(t) \qquad (1)$$

Differentiating both sides of equation 1 and assuming the system is initially at rest gives a linear ordinary differential equation with constant coefficients.

$$L \frac{di^2}{dt^2} + R \frac{di}{dt} + \frac{1}{C} i = v'(t) \qquad (2)$$

There are several techniques for solving equation 2 but Laplace methods are straightforward and give

$$Ls^2 I(s) + RsI(s) + \frac{1}{C} I(s) = sV(s) \qquad (3)$$

or

$$I(s) = \frac{\frac{1}{L} s \, V(s)}{s^2 + \frac{R}{L}s + \frac{1}{LC}} \qquad (4)$$

where s is the complex Laplace variable.

Example 1:

If L = 1h., C = 1f. and R = 2Ω in figure 1 find the time expression for the current if $v(t) = \mu(t)$ (the unit step function).

$$I(s) = \frac{1}{s^2 + 2s + 1} = \frac{1}{(s+1)^2}$$

$$i(t) = te^{-t}$$

An electrical engineer's interests are not restricted to analysis of electrical circuits. For example, non-electrical systems can be simulated by electrical circuits. One advantage of this simulation is the component values in the electrical circuit are relatively easily changed. Thus different non-electrical systems can often be compared more easily by comparing the corresponding electrical models. Also different forcing functions can be examined by simply changing the sources in the electrical circuit that models the system.

The basis for simulation stems from the similarity of the defining differential equations. Consider the mechanical system of figure 2. The defining equation for the system if the mass is initially at rest is

$$mx'' = -kx - bx' + f(t) \qquad (5)$$

or

$$mx'' + bx' + kx = f(t) \qquad (6)$$

Note the similarity of equations (2) and (6). This similarity points out the fact that the mechanical system of figure 2 can be "modeled" by the electrical circuit depicted in figure 1 if

$$\begin{aligned} L &= m \\ R &= b \\ C &= 1/k \end{aligned} \qquad (7)$$

SIMULATION WITH DISTRIBUTED NETWORKS

Not all systems to be simulated are described by linear ordinary differential equations. This paper deals with the simulation of those systems described by partial differential equations. First, the physical system will be defined and then described by a partial differential equation with appropriate boundary conditions. Next, the Laplace transform will be applied to the partial differential equation to obtain an expression for the solution. Finally, this expression will be synthesized as an electrical circuit with an excitation that will give the solution as a function of time.

In all cases the Laplace expression for the solution will contain terms like $e^{\sqrt{s}}$ or $\sqrt{s}$. Therefore the electrical circuit cannot be constructed with a finite number of resistors, capacitors and inductors alone but must include uniform distributed RC networks. To simplify the form of the Laplace expression for the solution and to show that the synthesis procedures are similar to those used in analog computer programming, the following substitutions will often be used:

$$\begin{aligned} W &= e^{\sqrt{s}} \\ \theta &= \sqrt{s} \end{aligned} \qquad (8)$$

Distributed devices are formed by placing a layer of dielectric material between layers of resistive and conductive materials as in figure 3. Uniformly-distributed RC devices (symbolized as in figure 4) have per unit resistances and per unit capacitances equal to constants r and c at any point along the length of the distributed device. The defining equations for the device are:

$$\begin{aligned} \frac{\partial V(x,s)}{\partial x} &= rI(x,s) \\ \frac{\partial I(x,s)}{\partial x} &= scV(x,s) \end{aligned} \qquad (9)$$

Considering a finite length d of the distributed element as a two port gives the solution

$$\begin{bmatrix} V(d,s) \\ I(d,s) \end{bmatrix} = e^{kd} \begin{bmatrix} V(o,s) \\ I(o,s) \end{bmatrix}$$

where

$$K = \begin{bmatrix} o & r \\ sc & o \end{bmatrix} \qquad (10)$$

This solution can be expressed in terms of hyperbolic functions of $\sqrt{sRC}$ and converted to the immitance formalisms

$$[Z] = \sqrt{\frac{R}{sc}} \begin{bmatrix} \coth\sqrt{RCs} & \operatorname{csch}\sqrt{RCs} \\ \operatorname{csch}\sqrt{RCs} & \coth\sqrt{RCs} \end{bmatrix} \qquad (11)$$

$$[Y] = \sqrt{\frac{sC}{R}} \begin{bmatrix} \coth\sqrt{RCs} & -\operatorname{csch}\sqrt{RCs} \\ -\operatorname{csch}\sqrt{RCs} & \coth\sqrt{RCs} \end{bmatrix} \qquad (12)$$

In the work that follows it will be necessary to consider cascaded uniformly-distributed RC elements. In figure 5 the ratio V_2/V_1 can be found as follows

$$T_{sc} = \frac{V_2}{V_1} = \frac{-y_{21_a}}{y_{22_a}+y_{11_b}} = \frac{\sqrt{\frac{sC_1}{R_1}} \operatorname{csch} \sqrt{sR_1C_1}}{\sqrt{\frac{sC_1}{R_1}} \coth \sqrt{sR_1C_1} + \sqrt{\frac{sC_2}{R_2}} \coth \sqrt{sR_2C_2}} \qquad (13)$$

Taking $C_1/R_1 = C_2/R_2$ gives

$$T_{sc}(s) = \frac{\sinh \theta_2}{\sinh(\theta_1+\theta_2)} \tag{14}$$

where $\theta_1 = \sqrt{R_1C_1s}$ and $\theta_2 = \sqrt{R_2C_2s}$.

In figure 6 the ratio V_2/V_1 is found to be

$$T_{oc} = \frac{V_2}{V_1} = \frac{-y_{21_a}}{y_{22_a} + \frac{1}{z_{11_b}}} \tag{15}$$

$$= \frac{\sqrt{\frac{sC_1}{R_1}}\, \operatorname{csch}\sqrt{sR_1C_1}}{\sqrt{\frac{sC_1}{R_1}}\, \coth\sqrt{sR_1C_1} + \sqrt{\frac{sC_2}{R_2}}\, \tanh\sqrt{sR_2C_2}}$$

Again taking $C_1/R_1 = C_2/R_2$ gives

$$T_{oc}(s) = \frac{\cosh \theta_2}{\cosh(\theta_1 + \theta_2)} \tag{16}$$

Now that these introductory comments on distributed networks have been made consider the following examples.

Example 2:

Find the temperature as a function of time at an arbitrary point x in a bar (figure 7) with diffusivity k, with insulated lateral surfaces and with its ends x = 0 and x = 1 kept at temperatures zero and F_o, respectively, when the initial temperature is zero everywhere except x = 1.

The boundary value problem is then

$$\begin{aligned} &u_t(x,t) = ku_{xx}(x,t) \qquad 0<x<1,\ t>0 \\ &u(x,0) = 0 \\ &u(0,t) = 0 \quad u(1,t) = F_o \end{aligned} \tag{17}$$

The transformed problem is

$$\begin{aligned} &su(x,s) = ku_{xx}(x,s) \\ &u(0,s) = 0 \quad u(1,s) = \frac{F_o}{s} \\ &u(x,0) = 0 \end{aligned} \tag{18}$$

This is then an ordinary differential equation with solution

$$u(x,s) = \frac{F_o}{s} \frac{\sinh x\sqrt{s/k}}{\sinh \sqrt{s/k}} \tag{19}$$

At $x = x_o$ proper choice of θ_1 and θ_2 in equation 14 will give an electrical model as in figure 5. If V_1 is a step voltage of magnitude F_o, then V_2 will be the desired temperature profile.

The desired solution of example 2 could have been found by using the definition of the inverse Laplace transform on equation 19. However, this solution is in the form of an infinite series which would necessitate the use of a digital computer. The use of a digital computer is not without convergence problems. In this problem it is usual to use

$$u(x_o,t) = F_o \sum_{n=0}^{\infty} \left[\operatorname{erf}\left(\frac{2n+1+x_o}{2\sqrt{kt}}\right) - \operatorname{erf}\left(\frac{2n+1-x_o}{2\sqrt{kt}}\right) \right] \tag{20}$$

for small values of t and

$$u(x_o,t) = F_o \left[x_o + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} e^{-n^2\pi^2kt} \sin n\pi x_o \right] \tag{21}$$

for large values of t. Furthermore, use of the digital computer requires additional equipment to obtain the solution as an analog waveform. These problems are eliminated when a distributed network model is used.

Example 3:

The initial temperature of a slab of homogeneous material with diffusivity k bounded by the planes $x = 0$ and $x = \ell$ is u_o. Find the temperature for all time at an arbitrary point x_o in the solid after the face $x = 0$ is insulated and the temperature of the face $x = \ell$ is reduced to zero (figure 8). The boundary value problem is then

$$\begin{aligned} &u_t(x,t) = ku_{xx}(x,t) \qquad 0<x<\ell,\ t>0 \\ &u(x,0) = u_o \\ &u_x(0,t) = 0 \qquad u(\ell,t) = 0 \end{aligned} \tag{22}$$

The transformed problem is then

$$\begin{aligned} &su(x,s) - u_o = ku_{xx}(x,s) \\ &u_x(0,s) = 0 \qquad u(\ell,s) = 0 \end{aligned} \tag{23}$$

which has the solution at x_o

$$u(x_o,s) = u_o \left[\frac{1}{s} - \frac{1}{s} \; \frac{\cosh x_o\sqrt{s/k}}{\cosh \ell\sqrt{s/k}} \right] \tag{24}$$

Equation 24 can be synthesized by the circuit of figure 9 where θ_1 and θ_2 are properly chosen and v(t) is a step of magnitude u_o.

It is not always easy to find the proper distributed network to solve the boundary value problem. However since the expressions we are considering are rational in W and θ (as described in equation 8) state variable synthesis methods can be used since the following basic blocks are available:

$$\frac{V_2}{V_1} = \frac{1}{W} = e^{-\sqrt{s}} \tag{25}$$

and

$$\frac{V_2}{V_1} = \frac{1}{\theta} = \frac{1}{\sqrt{s}} \tag{26}$$

The transfer function of equation 25 can be produced by choosing $\theta_2 >> \theta_1$ in the circuits of figures 5 or 6. For example, equation 14 gives

$$T_{sc}(s) = \frac{\sinh \theta_2}{\sinh(\theta_1+\theta_2)} = \frac{e^{-\theta_1}(1-e^{-2\theta_2})}{1-e^{-2(\theta_1+\theta_2)}} \tag{27}$$

But if $\theta_2 >> \theta_1$ then

$$|e^{-2\theta_2}| << 1$$
$$|e^{-2(\theta_1+\theta_2)}| << 1 \tag{28}$$

so

$$T_{sc}(s) \approx e^{-\theta_1} = \frac{1}{W} \tag{29}$$

The element of equation 26 can be produced by combining 1/W blocks. (see reference 1) With the introduction of these W-plane and θ-plane integrators many additional problems can be solved.

Example 4:

Find the temperature for all $t \geq 0$ at an arbitrary point x_o in a semi-infinite slab, $x > 0$, of homogeneous material under the condition of linear flow. The boundary $x = 0$ is kept at a constant temperature u_o and the initial temperature of the slab is zero throughout.

The boundary value problem is then

$$u_t(x,t) = ku_{xx}(x,t) \qquad x > 0 \quad t \geq 0$$
$$u(o,t) = u_o \qquad u(x,o) = 0 \tag{30}$$

The transformed problem is

$$su(x,s) = ku_{xx}(x,s)$$
$$u(o,s) = \frac{u_o}{s} \qquad u(x,o) = 0 \tag{31}$$

which has solution at x_o

$$u(x_o,s) = \frac{u_o}{s} e^{-x_o\sqrt{s/k}}$$
$$= \frac{u_o}{s} \left(\frac{1}{W}\right) \tag{32}$$

The solution is then found as the step response of a 1/W block. (a W-plane integrator)

Example 5:

Find the temperature for all time at an arbitrary point x_o in a semi-infinite solid under the condition of radiation at x = 0 into the solid at initial temperature u_o.

The boundary value problem is

$$u_t(x,t) = ku_{xx}(x,t) \qquad x > 0 \quad t > 0$$
$$u(x,o) = u_o \tag{33}$$
$$-u_x(o,t) + hu(o,t) = 0$$

The transformed problem is

$$ku_{xx}(x,s) - su(x,s) = -u_o$$
$$-u_x(o,s) + hu(o,s) = 0 \tag{34}$$

Solving equations 34 and using the W and θ substitutions gives at $x = x_o$

$$u(x_o,s) = \frac{u_o}{s}\left[1 - \frac{h}{(\theta+h)W}\right]$$

If v_1 is a step function of magnitude u_o in the circuit of figure 10 then v_2 will be the desired time solution.

CONCLUSION

It has been shown that distributed networks can be used to simulate the solution of certain types of partial differential equations. The examples used have been heat problems but the techniques apply equally well to various other areas such as the study of diffusion processes and analysis of seismic data. There appear to be applications in areas other than solution of partial differential equations as a final example exhibits.

Example 6:

Simulate the solution of the following linear, ordinary differential equation with variable coefficients.

$$2t \frac{dy}{dt} + y = 0$$

$$-2 \frac{d}{ds} (sy(s)) + y(s) = 0$$

$$-2(sy'(s) + y(s)) + y(s) = 0$$

$$-2sy'(s) - y(s) = 0$$

$$2s \frac{dy}{ds} + y = 0$$

$$\frac{dy}{y} = - \frac{ds}{2s}$$

$$\ell n\ y = - \frac{1}{2} \ell n\ s = \ell n \frac{1}{\sqrt{s}}$$

$$y(s) = \frac{1}{\sqrt{s}}$$

y(t) is then the impulse response of a θ-plane integrator.

REFERENCES

1. Hensley, R.D. and Bourquin, J.J., "Realizations of a W-Plane Integrator for Synthesis of Active Distributed Networks", 14th Midwest Symposium on Circuit Theory, Denver Colorado, May 7, 1971.
2. Hensley, R.D. and Bourquin, J.J., "Synthesis of Active Distributed RC Networks Via the W-Plane Integrator", 21st Southwestern I.E.E.E. Conference and Exhibition, Houston, Texas, April 30, 1971.

ACKNOWLEDGEMENT

The research for this paper was supported in part by NSF Grant GK 5335.

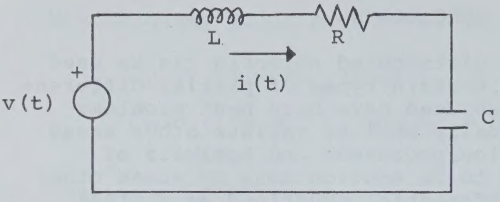


FIGURE 1

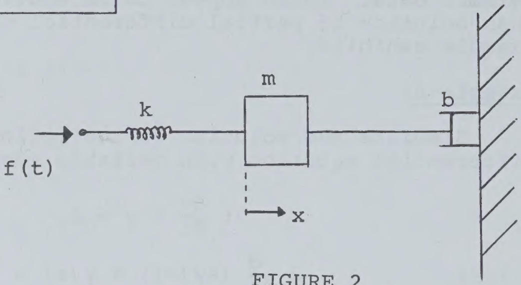


FIGURE 2

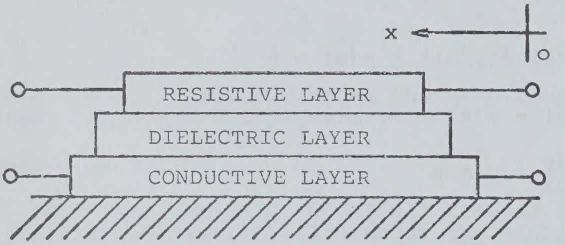

x
o
RESISTIVE LAYER
DIELECTRIC LAYER
CONDUCTIVE LAYER

FIGURE 3

SUBSTRATE

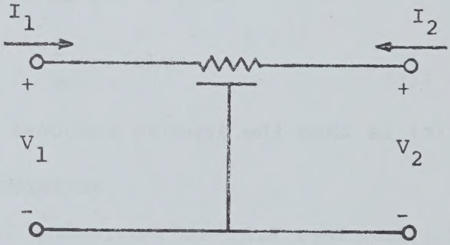


FIGURE 4

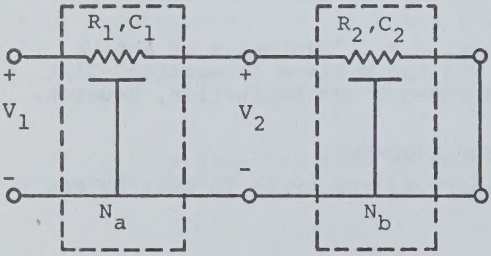

R1,C1
R2,C2
+
V1
-
+
V2
-
Na
Nb

FIGURE 5

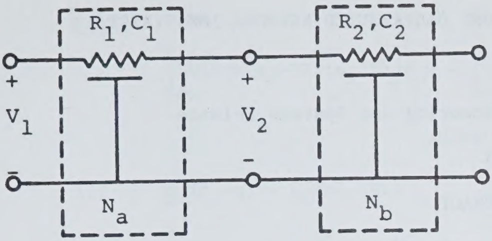

R_1, C_1
R_2, C_2
+
V_1
−
+
V_2
−
N_a
N_b

FIGURE 6

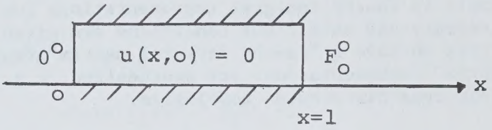

0^o
u(x,o) = 0
F^o_o
o
x
x=1

FIGURE 7

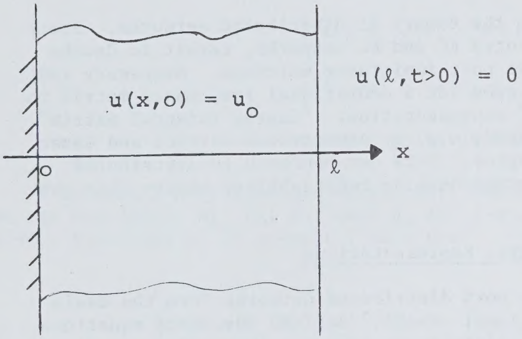

u(ℓ,t>0) = 0
$u(x,o) = u_o$
x
o
ℓ

FIGURE 8

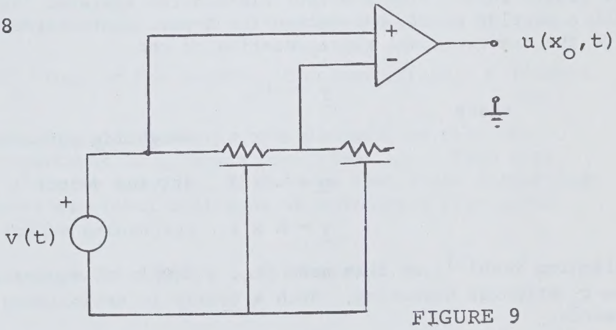


FIGURE 9

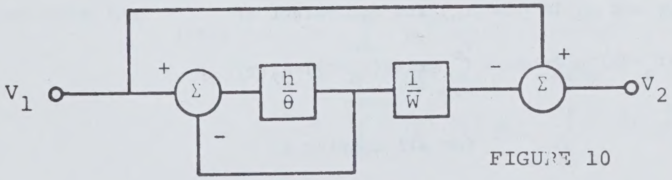

V_1
+
Σ
$\frac{h}{\theta}$
−
$\frac{1}{W}$
−
+
Σ
V_2

FIGURE 10

ON A NEW CHARACTERIZATION OF TWO PORT DISTRIBUTED NETWORK IMMITTANCE MATRICES

K.S. BAJAJ and J.A. RESH
Department of Electrical Engineering and Systems Science
Michigan State University
East Lansing, Michigan 48823

ABSTRACT

State equations for N-port distributed parameter networks are derived. State equations for two port distributed RC and RL networks, result in Cauchy integral representations for two port immittance matrices. Necessary and sufficient conditions are given for a nonrational immittance matrix to have a "Cauchy integral matrix" representation. "Cauchy integral matrix" representations are synthesized, e.g. by symmetrical lattice and Cauer type distributed topologies.

INTRODUCTION

This paper contributes to the theory of distributed networks. State equations, for two port distributed RC and RL networks, result in Cauchy integral representations for two port immittance matrices. Necessary and sufficient conditions are developed for a nonrational immittance matrix to have a "Cauchy integral matrix" representation. "Cauchy integral matrix" representations can be synthesized, e.g. by symmetrical lattice and Cauer type distributed network topologies. This new approach to distributed parameter networks opens up a comprehensive realizability theory like that for lumped networks.

1. Motivation for Cauchy Integral Representations

State descriptions of two port distributed networks form the basis for Cauchy integral representations. Resh(1) derived the state equations for single input, single output distributed systems. We will generalize, with a simpler proof, his method for N-port distributed networks. We start with the convolutional representation of the

$$\underset{\sim}{y} = z * \underset{\sim}{u}, \qquad (1)$$

where

$z = n \times n$ weighting pattern

$\underset{\sim}{u} = n \times 1$ driving vector

$\underset{\sim}{y} = n \times 1$ responding vector

Following Resh(1), we then seek Ht_o, a family of equivalence classes of pre-t_o stimulus histories. Such a family is established by the following theorem:

Theorem 1: $\underset{\sim}{u}_1$ and $\underset{\sim}{u}_2$ in $\underset{\sim}{u}(\infty,t_o]$ are equivalent if

$$\int_{-\infty}^{t_o} \exp(s(t_o-\tau))\cdot\underset{\sim}{u}_1(\tau)d\tau = \int_{-\infty}^{t_o} \exp(s(t_o-\tau))\cdot\underset{\sim}{u}_2(\tau)\cdot d\tau$$

for all complex s.

Proof:

$$\int_{-\infty}^{t_o} \exp(s(t_o-\tau))\cdot \underset{\sim}{u}_1(\tau)d\tau = -\int_{-\infty}^{o} e^{sx}\, \underset{\sim}{u}_1(t_o-x)dx$$

where $(t_o-\tau) = x$

let $\underset{\sim}{u}_1(t_o-x) = \underset{\sim}{f}_1(x)$ then

$$\int_{o}^{\infty} \underset{\sim}{f}_1(x)\, e^{sx} = \underset{\sim}{F}_1(-s)$$

similarly

$$\int_{o}^{\infty} \underset{\sim}{f}_2(x)\, e^{sx} = \underset{\sim}{F}_2(-s)$$

where $\underset{\sim}{f}_2(x) = \underset{\sim}{u}_2(t_o-x)$

now $\underset{\sim}{F}_1(-s) = \underset{\sim}{F}_2(-s)$ implies

$$\underset{\sim}{f}_1(x) = \underset{\sim}{f}_2(x) \quad \forall\ x > 0$$

or

$$\underset{\sim}{u}_1(t_o-x) = \underset{\sim}{u}_2(t_o-x) \quad \forall\ x > 0$$

or

$$\underset{\sim}{u}_1(t) = \underset{\sim}{u}_2(t) \quad \forall\ t \leq t_o$$

Identical histories are equivalent, hence the proof of the theorem. We can now define $H'_{t_o}[\underset{\sim}{u}_o]$ for each $\underset{\sim}{u}_o$ in $\underset{\sim}{u}(-\infty,t_o)$ to consist of those vector functions $\underset{\sim}{u}'_o$ in $\underset{\sim}{u}(-\infty,t_o]$ such that

$$\int_{-\infty}^{t_o} \exp(s(t_o-\tau))\cdot \underset{\sim}{u}'_o(\tau)d\tau = \int_{-\infty}^{t_o} \exp(s(t_o-\tau))\underset{\sim}{u}_o(\tau)d\tau$$

Next we want to find a state set $\Sigma(t_o)$ for the partition H'_{t_o} just defined.

A suitable lable for $H'_{t_o}(u_o)$ is the pattern of values $[\underset{\sim}{x}(s,t_o) = \int_{-\infty}^{t_o}\exp(s.(t_o-\tau)).\ \underset{\sim}{u}_o(\tau)d\tau]$ which characterizes its members. $\Sigma(t_o)$ itself is then the collection of all such patterns as $\underset{\sim}{u}_o$ ranges over $\underset{\sim}{u}(-\infty,t_o]$. With this identification of the state at any time, it follows that state transitions are governed by the two-dimensional continuum of ordinary differential equations given in Eq. (2):

$$\frac{d}{dt}\,\underset{\sim}{x}(s,t) = s\cdot\underset{\sim}{x}(s,t) + \underset{\sim}{u}(t) \qquad \forall \text{ Complex } s \tag{2}$$

For the relation between states and responses we return to Eq. (1) and note that

$$\begin{aligned}(z*\underset{\sim}{u})\ (t) &= \int_{-\infty}^{-\infty} [\underset{\sim}{u}^T(x)\ Z^T(t-x)dx]^T \\ &= \int_{-\infty}^{t} [\underset{\sim}{u}^T(x)\ Z^T(t-x)dx]^T \\ &= \int_{-\infty}^{t} [\underset{\sim}{u}^T(x)\ \frac{I}{2\pi i} \int_{\underset{\alpha}{-1}} Z^T(s)\ \epsilon^{s(t-x)}ds)dx]^T\end{aligned}$$

$$= \frac{1}{2\pi i} \int_{\mathcal{L}^{-1}} Z(s) \; \underset{\sim}{x} \; (s,t) \, ds$$

where $\mathcal{L}^{-1}$ is any Laplace inversion contour for Z(s).

Thus the system response is given by Eq. (3)

$$\underline{Y}(t) = \frac{1}{2\pi i} \int_{\mathcal{L}^{-1}} Z(s) \; \underline{x} \; (s,t) \tag{3}$$

In particular for two port case, the state equations are

$$\frac{d}{dt} \begin{bmatrix} x_1(s,t) \\ \\ x_2(s,t) \end{bmatrix} = \begin{bmatrix} s & 0 \\ \\ 0 & s \end{bmatrix} \begin{bmatrix} x_1(s,t) \\ \\ x_2(s,t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ \\ u_2(t) \end{bmatrix} \quad \forall \text{ complex } s \tag{4}$$

$$\begin{bmatrix} y_1(t) \\ \\ y2(t) \end{bmatrix} = \int_{\mathcal{L}^{-1}} \begin{bmatrix} Z_{11}(s) & Z_{12}(s) \\ \\ Z_{21}(s) & Z_{22}(s) \end{bmatrix} \begin{bmatrix} x_1(s,t) \\ \\ x_2(s,t) \end{bmatrix} ds \tag{5}$$

For RC and RL functions the line of discontinuity L (cf appendix) is negative realaxis. After evaluating the Laplace inversion integral and using the conditions to be stated later on, we get the state equations for such networks as

$$\frac{d}{dt} \begin{bmatrix} x_1(-x,t) \\ \\ x_2(-x,t) \end{bmatrix} = \begin{bmatrix} -x & 0 \\ \\ 0 & -x \end{bmatrix} \begin{bmatrix} x_1(-x,t) \\ \\ x_2(-x,t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ \\ u_2(t) \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} y_1(t) \\ \\ y_2(t) \end{bmatrix} = \int_L \begin{bmatrix} f_{11}(x) & f_{12}(x) \\ \\ f_{21}(x) & f_{22}(x) \end{bmatrix} \begin{bmatrix} x_1(-x,t) \\ \\ x_2(-x,t) \end{bmatrix} dx \tag{7}$$

Now if $u_1(t)$ and $u_2(t)$ are input currents and $y_1(t)$ and $y_2(t)$ are output voltages then the impedance matrix of two port RC distributed networks is given by

$$F(s) = \int_L \frac{\begin{bmatrix} f_{11}(x) & f_{12}(x) \\ \\ f_{21}(x) & f_{22}(x) \end{bmatrix}}{s + x} dx \tag{8}$$

Having observed that we can represent two port distributed network immittances by Eq. (8), we will develop a comprehensive theory for such representations, when the nonrational two-port immittance matrices are given.

2. Cauchy Integral Representations

The representation

$$F(s) = \frac{1}{2\pi i} \int_L \frac{[f(t)]dt}{t-s}$$

where $\frac{1}{2\pi i}[f(t)]$ is called the density function matrix, will be called Cauchy integral matrix representation for F(s).

Theorem 2: If a matrix F(s) has the following properties

B_1: $F(s) \rightarrow \underset{\sim}{0}$ (where $\underset{\sim}{0}$ is a null matrix) as $s \rightarrow \infty$

B_2: F(s) is sectionally holomorphic with line of discontinuity L

B_3: F(s) satisfies the boundary condition

$$\overset{-}{F}(t) - \overset{+}{F}(t) = 2\pi i \, f(-t)$$

where f(-t) is a matrix defined on L. Then F(s) is uniquely determined in the entire complex plane except $s \in L$.

Proof of this theorem is the generalization of the proof when f(s) is a scalar function, which is given in reference (3) on page 457.

Theorem 3: If a matrix F(s) satisfies B_1, B_2, B_3 and also the density function matrix belongs to class H* then the Cauchy integral matrix with density function matrix - [f(-t)] is the unique representation for F(s) i.e.

$$F(s) = \int_L \frac{-[f(-t)]dt}{t-s}$$

Conversely if F(s) has a Cauchy integral matrix representation with density function matrix satisfying H condition, then it satisfies B_1, B_2 and B_3.

Again the proof is a generalization of the proof given in reference (2).

3. Realizability Conditions

F(s) given in equation (8) can be realized if

I $f_{11}(x)$ and $f_{22}(x) > 0 \ \forall \ x$ on L

II $f_{12}(x)$ and $f_{21}(x)$ real

III $f_{11}(x) \cdot f_{22}(x) \geqslant f_{12}(x) \, f_{21}(x) \ \forall \ x$ on L

Conditions I, II, and III are of the same type as the lumped cases. In recognition of the similarities between Cauchy integrals and partial fractions we will refer to $f_{11}(x)$, $f_{12}(x)$, $f_{21}(x)$ and $f_{22}(x)$ as distributed residues.

4. Synthesis Procedures for Integral Representations of Two-port Networks

4.1 Symmetrical Lattice Networks: Symmetrical lattice networks as shown in fig. 1 have long been used in classical network theory (4). We will use them for realization of F(s) given in equation (8) and satisfying conditions I, II, and III. For symmetrical lattice networks

$$Z_a = Z_{11} - Z_{12} \tag{9}$$

$$Z_b = Z_{11} + Z_{12} \tag{10}$$

As $Z_{11} = Z_{22}$ so $f_{11}(x) = f_{22}(x)$ in eqn. (4.7)

$$Z_a = \int_L \frac{[f_{11}(x) - f_{12}(x)]dx}{s+x} \tag{11}$$

Now as $[f_{11}(x) - f_{12}(x)] \geqslant 0$ and if

$\int_L x^{(k-1)}(f_{11}(x) - f_{12}(x))dx$ exists for k=1,2, ...

then

$$Z_a^{[5]} = \frac{1|}{|b_1 s} + \frac{1|}{|b_2} + \frac{1|}{|b_3 s} + \frac{1|}{|b_4} + \ldots \tag{12}$$

$Z_b^{[5]} = \int_L \frac{[f_{11}(x) + f_{12}(x)]dx}{s+x}$ as $f_{11}(x)\ f_{12}(x) \geqslant 0$ and if

$\int_L x^{k-1}[f_{11}(x) + f_{12}(x)]dx$ exist then

$$Z_b = \frac{1|}{|c_1 s} + \frac{1|}{|c_2} + \frac{1|}{|c_3 s} + \frac{1|}{|c_4} + \ldots \tag{13}$$

Hence each arm of the lattice network is of the form shown in Fig. (2a and 2b).

Instead of having ladder networks for Z_a and Z_b we can also find Foster type of networks for them in fig. (3a). Continuum of networks as shown in fig. (3a) will be represented by fig. (3b).

4.2. Cauer's Networks: The network configuration used in Cauer's method of network realization [4] is shown in fig. (4). We will use this type of configuration for realization of Eqn. (8).

For network given in fig. (4)

$$Z_a = Z_{11} - \frac{1}{a} Z_{12} \tag{14}$$

$$Z_b = \frac{1}{a^2} \cdot Z_{22} - \frac{1}{a} Z_{12} \tag{15}$$

$$Z_c = \frac{1}{a} Z_{12} \qquad (16)$$

For F(s) given in eq. (8)

$$Z_a = \int_L \frac{[f_{11}(x) - \frac{1}{a} f_{12}(x)]dx}{s + x} \qquad (17)$$

$$Z_b = \int_L \frac{[\frac{1}{a^2} f_{22}(x) - \frac{1}{a} f_{12}(x)]dx}{s + x} \qquad (18)$$

$$Z_c = \frac{1}{a} \int_L \frac{f_{12}(x)}{s + x} dx \qquad (19)$$

Now it is possible to pick a in such a way as to make

(i) $\hat{f}(x) = [f_{11}(x) - \frac{1}{a} f_{12}(x)] > 0$

(ii) $\hat{f}_1(x) = [\frac{1}{a^2} f_{22}(x) - \frac{1}{a} f_{12}(x)] > 0$

(iii) $\hat{f}_2(x) = \frac{1}{a} f_{12}(x) > 0$

thus we can realize Z_a, Z_b, Z_c by Foster type of networks. If $\int_L x^{k-1} f(x)dx$ $\int_L x^{k-1} f_1(x)dx$ and $\int_L x^{k-1} f_2(x)dx$. exist for k=1,2,3 ... then we can Z_a, Z_b, Z_c by ladder networks [5]. Next we will suggest ways to realize impedance matrices, having density function matrices which do not satisfy the above realizability conditions.

Theorem 4: If a matrix F(s) satisfies conditons B_2 and B_3 and has a density function matrix

$$\begin{bmatrix} f_{11}(x) & f_{12}(x) \\ f_{21}(x) & f_{22}(x) \end{bmatrix}$$

where $f_{11}(x)$ and $f_{22}(x)$ are nonpositive ∀ x but satisfying the conditions $f_{11}(x) \cdot f_{22}(x) \geqslant f_{12}(x) \cdot f_{21}(x)$ ∀ x, then F(s) can be realized by RL networks if $\frac{1}{s}$ F(s) has a Cauchy integral matrix representation.

Proof: If $\frac{1}{s}F(s)$ has a Cauchy integral matrix representation then

$$F(s) = \int_L \frac{s \begin{bmatrix} \frac{f_{11}(x)}{-x} & \frac{f_{12}(x)}{-x} \\ \frac{f_{12}(x)}{-x} & \frac{f_{22}(x)}{-x} \end{bmatrix}}{s + x} dx$$

where $\frac{f_{11}(x)}{-x}$ and $\frac{f_{22}(x)}{-x}$ of the density function matrix are positive and satisfy the condition $\frac{f_{11}(x) f_{22}(x)}{x^2} \geqslant \frac{f_{12}(x) f_{21}(x)}{x^2}$ hence realizable by RL distributed networks.

Theorem 5: Let

$$F(s) = \int_L \frac{\begin{bmatrix} f_{11}(x) & f_{12}(x) \\ f_{21}(x) & f_{22}(x) \end{bmatrix}}{s + x} dx + \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

where R is a constant matrix and $f_{11}(x)$ and $f_{22}(x)$ are nonpositive functions but satisfy the condition $f_{11}(x)f_{22}(x) \geqslant f_{12}(x)f_{22}(x) \quad \forall\ x$ then F(s) can be realized by RL networks if

(i)
$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \int_L \begin{bmatrix} \frac{f_{11}(x)}{-x} & \frac{f_{12}(x)}{-x} \\ \frac{f_{21}(x)}{-x} & \frac{f_{22}(x)}{-x} \end{bmatrix} dx + \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

(ii) $\frac{f_{11}(x)}{x}, \quad \frac{f_{12}(x)}{x}, \quad \frac{f_{22}(x)}{x}, \quad \frac{f_{21}(x)}{x}$ are integrable.

(iii) $\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$ is P. R.

First matrix is realizable by RL networks while the second one is realizable by lumped resistive networks.

Next we will consider the case where the density function matrix is real, i. e.

$$F(s) = \int_L \frac{\begin{bmatrix} f_{11}(x) & f_{12}(x) \\ f_{21}(x) & f_{22}(x) \end{bmatrix}}{s + x} dx$$

where $f_{11}(x)$, $f_{12}(x)$, $f_{21}(x)$, $f_{22}(x)$ are real, spearating each function into positive and negative parts we can write

$$F(s) = \int_L \frac{\begin{bmatrix} \overset{+}{f}_{11}(x) + \overset{-}{f}_{11}(x) & \overset{+}{f}_{12}(x) + \overset{-}{f}_{12}(x) \\ \overset{+}{f}_{21}(x) + \overset{-}{f}_{21}(x) & \overset{+}{f}_{22}(x) + \overset{-}{f}_{22}(x) \end{bmatrix}}{s + x} dx$$

Then, the following two theorems give the realizability conditions for F(s).

Theorem 6: Let $F(\infty) = R$ and $F(s) = F^{+}(s) + F^{-}(s)$ where

$$F^{+}(s) = \int_L \frac{\begin{bmatrix} \overset{+}{f}_{11}(x) & \overset{+}{f}_{12}(x) \\ \overset{+}{f}_{21}(x) & \overset{+}{f}_{22}(x) \end{bmatrix}}{s + x} dx \text{ and } F^{-}(s) = \int_L \frac{\begin{bmatrix} \overset{-}{f}_{11}(x) & \overset{-}{f}_{12}(x) \\ \overset{-}{f}_{21}(x) & \overset{-}{f}_{22}(x) \end{bmatrix}}{s + x} dx$$

exist and $R=\hat{R}_1+\hat{R}_2$ where

$$\hat{R}_1 = \int_L \begin{bmatrix} \frac{f_{11}^-(x)}{-x}dx & \frac{f_{12}^-(x)}{-x}dx \\ \frac{f_{21}^-(x)}{-x}dx & \frac{f_{22}^-(x)}{-x}dx \end{bmatrix}$$

$$\hat{R}_2 = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

exist then F(s) has a network realization if

(i) $F^+(s)$ is distributed RC realizable
(ii) $[F^-(s)+\hat{R}_1]$ is distributed RL realizable
(iii) $\hat{R}_2$ is resistive realizable

Theorem 7: Let $F(\infty)=R$ and $F(s)=F_1(s)+F_2(s)$ where

$$F_1(s) = \int_L \frac{\begin{bmatrix} f_{11}^+(x) & f_{12}^-(x) \\ f_{21}^-(x) & f_{22}^+(x) \end{bmatrix}}{s+x} \text{ and } F_2(s) = \int_L \frac{\begin{bmatrix} f_{11}^-(x) & f_{12}^+(x) \\ f_{21}^+(x) & f_{22}^-(x) \end{bmatrix}}{s+x} dx$$

exist and $R=R_1+R_2$ where

$$R_1 = \int_L \begin{bmatrix} \frac{f_{11}^-(x)}{-x} & \frac{f_{12}^+(x)}{-x} \\ \frac{f_{21}^+(x)}{-x} & \frac{f_{22}^-(x)}{-x} \end{bmatrix} dx, \quad R_2 = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

exist then F(s) has a network realization if

(i) $F_1(s)$ is distributed RC realizable
(ii) $[F_2(s)+R_1]$ is distributed RL realizable
(iii) R_2 is resistive realizable

Example 1: Let

$$F(s) = \begin{bmatrix} a_1 s^{-1/3} & -s^{1/2} \\ -s^{1/2} & a_2 s^{-1/3} \end{bmatrix}$$

As F(s) satisfies the conditions of theorem (3)
Hence the Cauchy integral representation is

$$F(s) = \int_0^\infty \frac{\begin{bmatrix} \frac{\sqrt{3}}{2\pi} a_1 x^{-1/3} & -\frac{x^{-1/2}}{\pi} \\ -\frac{x^{-1/2}}{\pi} & \frac{\sqrt{3}}{2\pi} a_2 x^{-1/3} \end{bmatrix}}{s+x} dx$$

For $a_1 = a_2$ if we have

$$\text{(i)} \quad Z_a = \int_0^\infty \left[\frac{\sqrt{3}}{2\pi} a_1 x^{-1/3} - \frac{x^{-1/2}}{\pi} \right] \geq 0$$

$$\text{(ii)} \quad Z_b = \int_0^\infty \left[\frac{\sqrt{3}}{2\pi} a_1 x^{-1/3} - \frac{x^{-1/2}}{\pi} \right] \geq 0$$

Then F(s) can be realized by Lattice networks. Realization of Z_a and Z_b is shown in fig. (3) where $f(x) = \frac{1}{\pi}. \left[\frac{\sqrt{3}}{2} a_1 x^{-1/3} - x^{-1/2} \right]$ and $f_1(x) = \frac{1}{\pi} \left[\frac{\sqrt{3}}{2} a_1 x^{-1/3} + x^{-1/2} \right]$.

F(s) can be realized by Cauer's method taking the turns ratio for the transformer to be -a.

Example 2:

$$F(s) = \begin{bmatrix} \left[5-2s \log \frac{s+2}{s+1} + 2 \log 2 \frac{(s+1)}{s+2}\right] & \left[\log \frac{s+1}{s+2} - 1 - s \log \frac{s+1}{s+2}\right] \\ \left[\log \frac{s+1}{s+2} - 1 - s \log \frac{s+1}{s+2}\right] & \left[10-2 s \log \frac{(s+2)}{s+1} + 2 \log \frac{s+1}{s+2}\right] \end{bmatrix}$$

which can be written as

$$F(s) = \begin{bmatrix} \int_1^2 \frac{2xdx}{s+x} & \int_1^2 - \frac{xdx}{s+x} \\ \int_1^2 - \frac{xdx}{s+x} & \int_1^2 \frac{2xdx}{s+x} \end{bmatrix} + \begin{bmatrix} s. \int_1^2 \frac{2/xdx}{s+x} & s. \int_1^2 \frac{1/xdx}{s+x} \\ \int_1^2 \frac{1/xdx}{s+x} & s. \int_1^2 \frac{2/xdx}{s+x} \end{bmatrix} + \begin{bmatrix} 3 & -\log 2 \\ -\log 2 & 8-2 \log 2 \end{bmatrix}$$

$$= F_1(s) + f_2(s) + f_3(s)$$

we can realize $f_1(s)$ by RC symmetrical lattice networks, having

$$Z_a = \int_1^2 \frac{3xdx}{s+x} \qquad Z_b = \int_1^2 \frac{xdx}{s+x}$$

$f_2(s)$ can be realized by RL symmetrical lattice networks having

$$Z_a^1 = s \int_1^2 \frac{1/xdx}{s+x} \qquad Z_b^1 = s \int_1^2 \frac{\frac{3}{x} dx}{s+x}$$

An algorithm used to get ladder networks for Z_a, Z_b, Z_a^1, Z_b^1 is given in reference (6). Ladder networks for Z_a, Z_b, Z_a^1, Z_b^1 are given in fig.(5).

Finally we can realize $f_3(s)$ by Cauer's method, by taking a= -1 hence the final network is obtained by connection the three realizations in series.

REFERENCES

1. J.A. Resh and I.C. Goknar "Derivation of Canonical State Equations for a Class of Distributed Systems" Sixth Allerton Conference on Circuit and System Theory, 1968.

2. N.I. Muskhelishuili, Singular Integral Equations (Translated from Russian Edition by J.R.M. Radok) Groningen, Holland: P. Noordhoff N.U. 1953.

3. Pogarzelski, Integral Equations and Their Applications. Volume 1, Pergamon Press.

4. VanValkenburg, Introduction to Modern Network Synthesis, John Wiley and Sons, New York 1967.

5. O. Perron, Die Lehre Von den Kettenbruchen Band II. B.G. Teubner Verlags Gessellshaft. Stuttgart 1957.

6. K. S. Bajaj, Cauchy Type Integral Representations for Network Functions, Ph.D. Thesis, Michigan State University 1971.

7. G.T. Daryanani, Foster Distributed Lumped Network Synthesis, Ph.D. Thesis, Michigan State University.

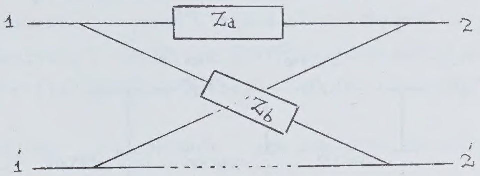


Fig (1)

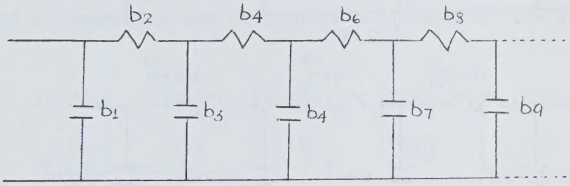

b2
b4
b6
b8
b1
b3
b4
b7
b9

Fig (2a)

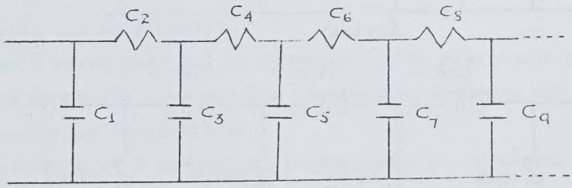

C2
C4
C6
C8
C1
C3
C5
C7
C9

Fig (2b)

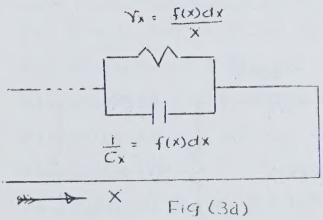

Yx = f(x)dx / x
1/Cx = f(x)dx
X

Fig (3a)

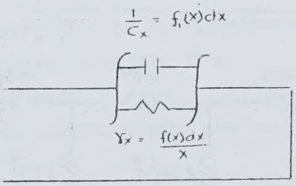

1/Cx = f1(x)dx
Yx = f(x)dx / x

Fig (3b)

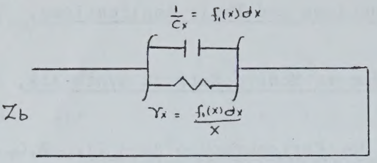

$\frac{1}{C_x} = f_1(x)dx$
Zb
$r_x = \frac{f_1(x)dx}{x}$

Fig (3C)

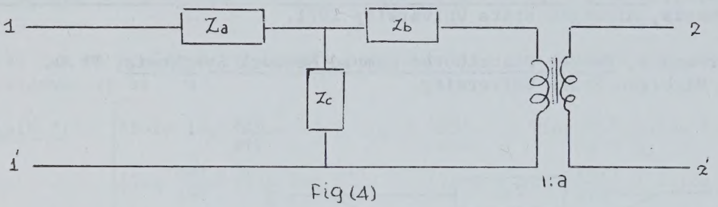

1
Za
Zb
2
Zc
1'
2'
1:a

Fig (4)

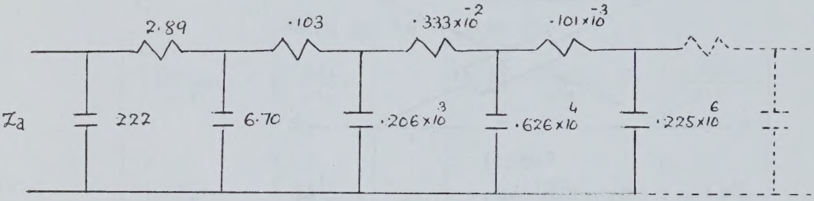

2.89
.103
$.333\times10^{-2}$
$.101\times10^{-3}$
Za
222
6.70
$.206\times10^{3}$
$.626\times10^{4}$
$.225\times10^{6}$

Fig (5a)

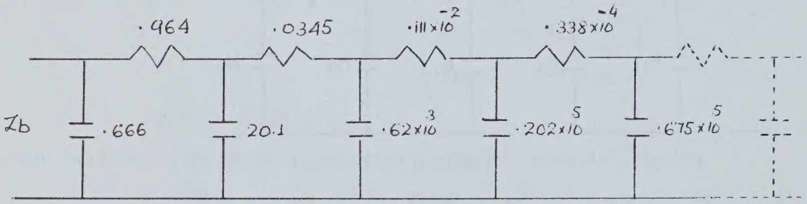

.964
.0345
$.111\times10^{-2}$
$.338\times10^{-4}$
Zb
.666
20.1
$.62\times10^{3}$
$.202\times10^{5}$
$.675\times10^{5}$

Fig (5b)

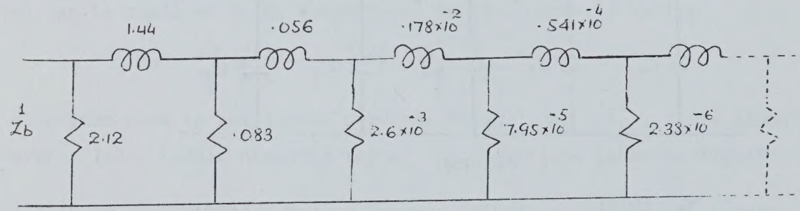
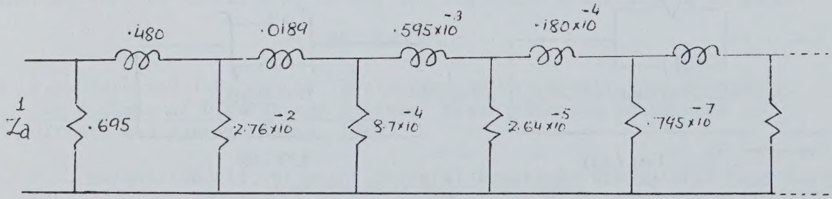

1.44
.056
$.178\times10^{-2}$
$.541\times10^{-4}$
1
Zb
2.12
.083
2.6×10^{-3}
7.95×10^{-5}
2.38×10^{-6}

Fig (5C)


.480
.0189
$.595\times10^{-3}$
$.180\times10^{-4}$
1
Za
.695
2.76×10^{-2}
8.7×10^{-4}
2.64×10^{-5}
$.745\times10^{-7}$

PROPERTIES OF NETWORK PARAMETERS OF COUPLED NONUNIFORM LC LINES

KISHORE SINGHAL AND JIRI VLACH
Department of Electrical Engineering
University of Waterloo, Waterloo, Ontario, Canada

The results presented here are an extension to those of Protonotarios and Wing [1] on single nonuniform LC lines.

The equations of a system of coupled nonuniform LC lines of length ℓ are given as:

$$\overset{\circ}{V} = -sL(x)J \quad ; \quad \overset{\circ}{J} = -sC(x)V$$

with L and C being n x n symmetric matrices. Furthermore we restrict the system so that:

a) All elements of L and C are analytic functions of x.

b) The matrices L and C are positive semi-definite at all points x.

c) L and C are positive definite 'in the average', i.e.

$$\int_0^\ell L(x)\,dx, \qquad \int_0^\ell C(x)\,dx \quad \text{are positive definite.}$$

d) The eigenvalues $\lambda_i(x)$ of LC are distinct for all x.

If the transmission and immittance parameters of the line (as a 2n-port) are

$$\begin{bmatrix} V(0) \\ J(0) \end{bmatrix} = \underbrace{\begin{bmatrix} \mathcal{A}(s) & \mathcal{B}(s) \\ \mathcal{C}(s) & \mathcal{D}(s) \end{bmatrix}}_{T(s)} \begin{bmatrix} V(\ell) \\ J(\ell) \end{bmatrix} \quad ; \quad \begin{bmatrix} V(0) \\ V(\ell) \end{bmatrix} = \underbrace{\begin{bmatrix} Z_{11}(s) & Z_{12}(s) \\ Z_{21}(s) & Z_{22}(s) \end{bmatrix}}_{Z(s)} \begin{bmatrix} J(0) \\ -J(\ell) \end{bmatrix}$$

and $Q = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$ (I being an identity matrix of appropriate order) then the following can be proved:

1. The $\mathcal{ABCD}$ parameters can be expressed as an everywhere convergent matrix series in powers of s and the coefficient matrices can be determined by successive integrations.
2. All elements of T are entire functions of s. Elements of $\mathcal{A}$ and $\mathcal{D}$ are even while those of $\mathcal{B}$ and $\mathcal{C}$ are odd.
3. $T(0) = I$, $\det(T) = 1$ and $Q\overset{\circ}{T}(0)$ is positive definite.
4. For $s = j\omega$ $T^*QT = Q$ and for Re $s>o$, T^*QT-Q is positive definite.
5. All elements of Z are ratios of entire functions of s.
6. All poles of Z lie on the imaginary axis.
7. All zeros of Z_{11} and Z_{22} lie on the imaginary axis and interlace with the poles.
8. All poles of Z and the zeros of the elements of Z_{11} and Z_{22} satisfy the asymptotic estimate

$$\lim_{r\to\infty} \frac{n(r)}{r} = \frac{2}{\pi} \sum_{i=1}^{n} \int_{0}^{\ell} \sqrt{\lambda_i(x)}\, dx = \Delta$$

n(r) being the number of poles or zeros with magnitude less than r.

9. The denominator functions of the elements of Z are of order 1, genus 1 and type $\Delta\pi/2$. The numerator functions of the diagonal terms of Z satisfy the same relations.
10. The k^{th} poles of Z and the corresponding zeros of the elements of Z_{11} and Z_{22} are located asymptotically at $\frac{2k}{\Delta}$.

Most of these results can be proved under less restrictive conditions. Similar results hold for multilayer RC nonuniform networks.

REFERENCE

1. E. N. Protonotarios and O. Wing, "Analysis and intrinsic properties of the general non-uniform transmission line", IEEE Trans. on Microwave Theory and Techniques, Vol. MTT-15, pp. 142-150, 1967.

DISCRETE DISTRIBUTED SYSTEMS

G. L. KUSIC, Department of Electrical Engineering, University of Pittsburgh, Pittsburgh, Pa. 15213

V. ARANDJELOVIC, "Boris Kidric" Institute of Nuclear Sciences, Belgrade

ABSTRACT

In this paper distributed parameter systems are examined from an approach analogous to lumped parameter systems. The distributed parameter system is assumed to be linear and constant coefficient, in order that properties of the Laplace and Fourier transforms such as superposition, etc. are valid. It is intended to develop discrete approximations to distributed system. The discrete version may then be used to calculate the response to forcing functions and initial conditions. Stability properties of the discrete equation are shown to be more involved than for ordinary differential equations. A general method is presented to select stable difference approximations to the partial differential equation. Many of the concepts used in this paper are extensions of aspects from linear control theory.

A - The Partial Differential Equation Impulse Response

Consider the N^{th} order in time, constant coefficient partial differential equation describing a distributed parameter system:

$$-\sum_{i=0}^{N} \alpha_i' \frac{\partial^i y}{\partial t^i}(t,x) = -\sum_{i=0}^{N}\sum_{k=0}^{M} b_{ik}\frac{\partial^k}{\partial x^k}\frac{\partial^i y}{\partial t^i} = f(t,x), b_{MN}=-1 \qquad \text{(A1)}$$

where x may be either a scalar or a vector defined on the domain:

$$R: \{0 \leq x \leq 1,\ 0 \leq t \leq t_{max}\} \qquad \text{(A2)}$$

and the variable t represents time.

The α_i' are partial differential operators with respect to the space axis, x. For example the equation

$$\frac{\partial^2 y}{\partial t^2} + 2\frac{\partial^2 y}{\partial x \partial t} + \frac{\partial^2 y}{\partial x^2} = f(t,x) \qquad \text{(A1.1)}$$

has the following definitions for $i \leq 2$, $k \leq 2$:

$b_{22} = -1$, $b_{21} = b_{20} = 0$, $b_{12} = 0$, $b_{11} = -2$, $b_{10} = 0$, $b_{02} = -1$

$b_{01} = b_{00} = 0$, $\alpha_2 = -1$, $\alpha_1 = -2\frac{\partial}{\partial x}$, $\alpha_0 = -\frac{\partial^2}{\partial x^2}$

The N initial conditions for the system are known for the range of x. In some cases a combination of initial and final values are known. Notice that in equation (1), a total of M boundary conditions must be given along x = 0 and x = 1. Stated in this way, the system described by eq. (1) is a mixed initial and boundary value problem. In the case where the boundary conditions are constant the distributed system may be reduced to a pure initial-value problem, (with forcing function f(t,x)), which has a block diagram similar to a lumped system:

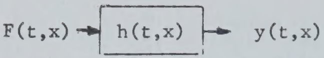


Figure 1 - Block Diagram for a Distributed System with Constant Boundary Functions

In fig. (1), h(t,x) represents the impulse response of the distributed system with all initial conditions set to zero. In addition, if the boundary conditions are not constant, these are additional inputs into the system. However, as the system is linear, stability properties depend solely upon the impulse response, h(t,x), and accordingly attention will be directed towards the impulse response.

Let the boundary conditions of eq. (A1) be set to zero, and the forcing function f(t,x) be repeated on the entire spatial axis - in order to generate a periodic function, y(t,x)=y(t,x+1) for 1 = ..., -3,-2,-1,0,1,2,....... In this situation the initial conditions and forcing functions represent periodic data for which it is sufficient to examine only the response for $0 \le X \le 1$. The periodic data are assumed to be sufficiently smooth and they may be described by a convergent Fourier series:

$$F(t,x)=\sum_{p=0}^{\infty} A_p(t)\,\varepsilon^{j2\pi p} \qquad \frac{\partial^i y}{\partial t^i}(0,x) = \sum_{p=0}^{\infty} C_{ip}\,\varepsilon^{j2\pi p} \qquad (A3)$$

When the data is expressed in the form of eq. (A3), and the solution uniquely depends on the data, it is evident that the partial differential equation, eq. (A1), carries each frequency component p into the solution. In other words, the solution to eq. (A1) may be expressed in terms of a Fourier series expansion. In order to examine the response of the distributed system to each frequency component, the complex Fourier transform /1/ of eq. (A1) is taken with respect to the space axis:

$$F_x\{-\sum_{i=0}^{N}\alpha'_i \frac{\partial^i y}{\partial t^i}(t,x)\} = -\int_{-\infty}^{+\infty}\varepsilon^{-jpx}\sum_{i=0}^{N}\alpha'_i\frac{\partial^i y}{\partial t^i}\,dx$$

$$= -\sum_{i=0}^{N}\alpha_i \frac{\partial^i Y}{\partial t^i}(t,p) = -\sum_{i=0}^{N}\sum_{k=0}^{M} b_{ik}(jp)^k\frac{\partial^i Y}{\partial t^i}(t,p) = F(t,p) \qquad (A4)$$

where jp is the dual or Fourier variable corresponding to $\partial/\partial x$. As an example of the transformation, consider the heat equation:

$$F_x\{\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial_x{}^2}\} = \frac{\partial U}{\partial t} + p^2 U = F(t,p) \qquad (A4.1)$$

where it is evident that the amplitude of each frequency component of the initial data (eq. (A3)) may be introduced for each p and the response found by solving the first order differential equations.

Equation (A4) is a complex coefficient polynomial analogous to the differential equation which describes a linear system. To continue the analogy, take the Laplace transform of equation (A4) in order to obtain the transfer function for the distributed parameter system (without repeated poles):

$$H(s,p) = \frac{Y(s,p)}{F(s,p)} = \frac{i}{-\sum_{i=0}^{N}\alpha_i s^i} = \sum_{i=1}^{N}\frac{r_i(p)}{S+\lambda_i(p)} \qquad (A5)$$

where r_i and λ_i are functions of the space frequency, p. Except in rare equations, the poles of eq. (A5) are neither integer multiples nor real functions of p. For the distributed system given by eqs. (A1), (A4), and (A5), the following definition is used /2/:

Definition 1 - A linear, constant coefficient, N^{th} order in time, distributed parameter system is said to be asymptotically stable in time if the poles of its transfer function (eq. (5)) are such that

$$\text{Real } \{\lambda_i(p)\} > 0$$

for i = 1,2,...N and all values of space frequency, p.

Stated as definition (1), a stable distributed system reaches a steady condition as time approaches infinity. The heat equation given by eq. (4a) has the transfer function and impulse response given by:

$$H_n(s,p) = \frac{1}{s + p^2} \qquad h(t,x) = \frac{\varepsilon^{-x^2/4t}}{2\sqrt{\pi t}} \tag{A4.2}$$

where the transfer function indicates that all frequency components reach a steady-state condition. It may equivalently be seen that $h(t,x) \to 0$ as $t \to \infty$.

B - Parabolic, Hyperbolic and Elliptic Systems

The traditional way to classify partial differential equations has been in terms of the roots of a second order equation /3/4/. These types are called parabolic, and elliptic if the roots of the second order in both time and space equation:

$$L_t F_x \{ \frac{\partial^2 u}{\partial t^2} + a \frac{\partial^2 u}{\partial x \partial t} + b \frac{\partial^2 u}{\partial x^2} \} = s^2 + asjp - bp^2 \tag{B1}$$

$$s = [-\frac{a}{2} \pm \frac{1}{2} \sqrt{a^2 - 4b}\ ]\ (jp) \tag{B2}$$

are respectively equal (a^2=4b), imaginary (a^2> 4b), or complex (a^2<4b). It is readily seen that for eq. (B2) the roots may be expressed as (s/jp), hence there is a proportionality between the x and t axes which define the "characteristics." The characteristics are given by the solution to the differential equation

$$(\frac{\partial t}{\partial x})^2 + a(\frac{\partial t}{\partial x}) + b = 0 \tag{B3}$$

For the parabolic equation, there is only one characteristic, but for the hyperbolic case (a^2>4b), there are two characteristics which define the domain of dependence of the solution. For example for when two initial conditions given for $0 \leq x \leq 1$, the solution is totally known for the region shown in figure (2):

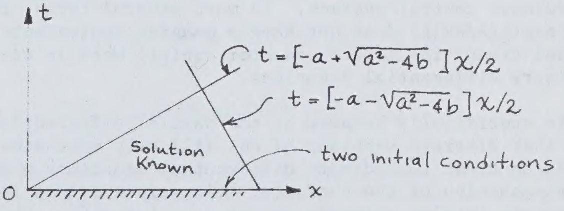


Figure 2 -Characteristics for a Second Order Partial Differential Equation

Various other combinations of boundary conditions (e. g. specifying the value of y along two intersecting characteristics) are possible for the hyperbolic case.

The mathematics literature is extremely rich in analyzing parabolic, hyperbolic, and elliptic equations /4/, /16/ and it will not be attempted to condense this material here. However, the significance of the roots of eq. (B1) from the mathematical point of view is that the roots determine required boundary conditons which, along with the equation itself, constitute a well-posed problem. For example, when the roots are complex, eq. (B1) would be elliptic and the boundary conditions should be either Neumann or Dirichelet types for a closed controur which encloses the domain of the problem. Accordingly, many properties of the solution are a-priori known such as the solution takes on maxima only on the boundaries for an elliptic equation.

Classifying an equation in terms of elliptic, hyperbolic, or parabolic is of doubtful usefulness for more than the restricted equation (B1) as the type of equation may change, say from parabolic to hyperbolic if $\hat{a}$ and b are functions of x and t. Even if this type of change occurs (as in fluid flow for example), the equation continues to describe a physically realistic process. Another point to consider is that it is possible to computationally satisfy the partial differential equation using a set of unusual boundary conditions. A notable example of such a case is the electron gun design methods of Kirstein[5] and Harker[6] in which the function and its normal derivative were specified along a curve, then Laplace's equation (i.e. the elliptic equation $\partial^2u/\partial x^2+\partial^2u/\partial t^2$) was solved inside a domain with unspecified terminal conditions by means of using the analytic continuation technique of Garabedian[5]. This is in essence formulating an elliptic equation as an initial-value problem, but the method required that a third dimension be utilized (analytic continuation of the initial data into the third dimension) in order to solve the problem.

C - Discrete Approximations

To continue the analogy of the distributed system with the single independent variable case, consider a time direction which is discretized while the space direction is continuous. In this situation, the Z-transform of the impulse response is given by

$$Z_t\{H(s,p)\} = Z_t\{\sum_{i=1}^{N} \frac{r_i(p)}{s+\lambda_i(p)}\} = \sum_{i=1}^{N} \frac{Zr_i(p)}{Z-\varepsilon^{-\lambda_i(p)T}} \qquad \text{(c1)}$$

where T is the uniform time increment and $Z = \varepsilon^{sT}$ is the unit advance operator. Notice that the space direction remains in terms of the dual variable p. This represents the source of difficulty when the space-time difference equation must be found for eq. (c1), because the quantity $\exp(-\lambda_i(p)T)$ does not have the conceptual equivalent of the exponential decay exp(aT) from linear ordinary control systems. In more general terms, the differential operator $\exp(\underline{A}(\partial/\partial x)t)$ does not have a meaning equivalent to the constant coefficient transition matrix operator $\exp(\underline{A}t)$ used in control theory[7] with ordinary differential equations.

It is specifically because of the partial differential exponential operator that discrete versions of eq. (A1) only approximate the solution. This is in contrast to ordinary differential equations wherein, if the Heavyside expansion of the characteristic polynomial is found, then the exact discrete impulse response may be calculated[8]. Therefore, the problem is to formulate a stable difference scheme in both time and space for eq. (Al), whose transfer approaches eq. (cl). The conditions which the difference approximation to (Al) must satisfy for the initial-value problem are well known and summarized by Von Neumann's Theorem/9/ - A necessary condition for stability of the difference scheme is the eigenvalues of the amplification matrix G(T,p) must be

$$|Z_i| \leq 1 + 0(T) \qquad \text{for} \begin{cases} 0 < T < T_{max} \\ \text{all } p \\ i = 1,2,.... \end{cases}$$

In this theorem, the space increment is Δx, T is the time increment. The space direction is Fourier transformed, and the space increment is assumed to be a function, $\Delta x = f(T)$ such that $f(0) = 0$. The matrix G(T,p) then represents a discrete transfer function which yields the solution for the next row in time in terms of the previous row. In essence this describes the feedforward element of the transfer function equivalent of figure 1. The structure is shown below:

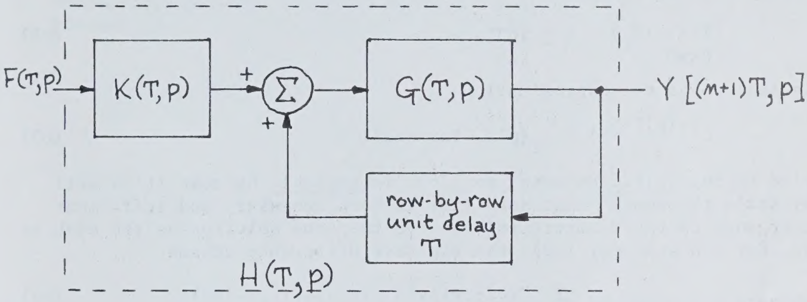


Figure 3. - Transfer Dunction for the Discrete Approximation to h(t,x) [where it is assumed that K(T,p) either does not have poles or that they may be cancelled by an appropriate choice of input function.]

As an example of the numerical process, consider the wave equation with f(t,x) = 0:

$$\frac{\partial^2 y}{\partial t^2}(t,x) = \frac{\partial^2 y}{\partial x^2}(t,x) \overset{F_x}{\longleftrightarrow} \frac{\partial^2 y}{\partial t^2} = -p^2 y(t,p) \tag{c2}$$

which may be approximated by the difference scheme

$$y(t+T,x)-2y(t,x)+y(t-T,x) \overset{\sim}{=} \left(\frac{T}{\Delta x}\right)^2 (y(t-T,x+\Delta x)-2y(t-T,x)+y(t-T,x-\Delta x)) \tag{c3}$$

If this equation is Fourier transformed with respect to the space axis, it becomes

$$Y(t+T,p)-2Y(t,p)+Y(t-T,p) \overset{\sim}{=} \left(\frac{T}{\Delta x}\right)^2 (\varepsilon^{jp\Delta x}-2+\varepsilon^{-jp\Delta x})Y(t-T,p) \tag{c4}$$

Since this last equation contains more than one time-step, let the new variable V(t,p)=Y(t-T,p) be introduced so that eq. (c4) may be brought into the form

$$\begin{bmatrix} V(t+T,p) \\ Y(t+T,p) \end{bmatrix} = \left[\begin{array}{c|c} 2 & -1+(\frac{T}{\Delta x})^2(-2+2\cos(p\Delta x)) \\ \hline 1 & 0 \end{array}\right] \begin{bmatrix} V(t,p) \\ Y(t,p) \end{bmatrix} = G(T,\Delta x,p) \begin{bmatrix} V(t,p) \\ Y(t,p) \end{bmatrix} \quad (c5)$$

where the space increment Δx is explicitly retained. This process of introducing dummy variables, $(V(t,p))$, may be carried out for any order in time equation. The eigenvalues of eq. (c5) may be easily found from its characteristic polynomial

$$p(Z) = Z^2-2Z+1+2(\frac{T}{\Delta x})^2(1-\cos(p\Delta x)) \quad (c6)$$

which could of immediately been written by inspection of eq. (c4). The roots of eq. (c6) are

$$Z_i = 1 \pm j(\frac{T}{\Delta x})\sqrt{2}\sqrt{1-\cos(p\Delta x)} \quad (c7)$$

which by Von Neumann's theorem is clearly unstable for any $T > 0$ and any frequency $p > 0$. However, as $\Delta x \to 0$

$$\lim_{\Delta x \to 0} (Z_i) = 1 \pm jpT \quad (c8)$$

which is seen to be the approximation

$$\varepsilon^{\lambda_i(p)T} \cong 1 + \lambda_i(p)T, \text{ for small } T \quad (c9)$$

as applied to eq. (c1). However, eq. (c3) is unusable because it is well known by Lax's theorem/9/ that stability is both necessary and sufficient for convergence of the discrete solution to the true solution as the mesh is refined. For the wave eq. (c2), the standard difference scheme

$$y(t+T,x)-2y(t,x)+y(t-T,x) \cong (\frac{T}{\Delta x})^2(y(t,x+\Delta x)-2y(t,x)+y(t,x-\Delta x)) \quad (c9)$$

which is a slight modification of eq. (c3), has the following roots for its characteristic polynomial:

$$Z_i = 1-(\frac{T}{\Delta x})^2(1-\cos(p\Delta x)) \pm j\sqrt{2(\frac{T}{\Delta x})^2(1-\cos p\Delta x)-(\frac{T}{\Delta x})^4(1-\cos p\Delta x)^2} \quad (c10)$$

Both roots of eq. (c10) lie inside the unit circle for $(\frac{T}{\Delta x}) < 1$ and for all $p > 0$, hence for these conditions eq. (c9) represents a stable difference scheme. For $\Delta x \to 0$ it may be verified that the roots given by eq. (c10) also approach $1 \pm jpT$, which again implies that the discrete eq. (c9) is consistent to the exact solution given by eq. (c1).

It is now evident that a discrete approximation to a partial differential equation may be consistent to the equation but it also must be stable in order to be a useful method. A general method to formulate the discrete approximation is given in the following sections.

D - Space Derivative Approximations

It is subsequently shown that when difference equations are used to obtain the discrete solution by means of a row-by-row calculation in time, the approximations to the space derivatives must have a "definiteness", *i.e.*

have negative and positive definite real parts[7] respectively for their Fourier transforms.

E - Time Derivative Approximations

Let the α'_i differential operators of eq. (A1) be approximated by difference expressions and the resulting differential-difference equation be Fourier transformed with respect to the space axis to yield:

$$\frac{\partial^N Y}{\partial t^N} - \alpha_{N-1}\frac{\partial^{N-1} Y}{\partial t^{N-1}} \cdots\cdots - \alpha_o Y(t,p) - F(t,p) \qquad \text{(E1)}$$

Next, the time derivatives of eq. (E1) may be approximated by difference expressions. There are a large variety of ways in which this may be accomplished. Since a change in variables will permit a difference scheme for a gth derivative to be reduced to a g + 1 step process, there is no loss in generality by expressing $\partial^N Y/\partial t^N$ of eq. (E1) by an N + 1 step approximation and expressing i = 0,1,2,...,N - 1 by means of a q + 1 step process where $q \geq N$. Hence eq. (E1) becomes:

$$\begin{array}{l} \frac{1}{T^N}\{\frac{Y^q}{0!} - \frac{NY^{q-1}}{1!} + \frac{N(N-1)Y^{q-2}}{2!} +++++++ (-1)^N Y^{q-N}\} \\ \\ \frac{-\alpha_{N-1}}{T^{N-1}}\{b_o Y^q + b_1 Y^{q-1} + + + b_N Y^{q-N} + b_{N-1} Y^{q-N-1} + + + b_2 Y^0\} \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ -\alpha_o\{d_o Y^q + + + + \qquad d_N Y^{q-N} + + + + d_q Y^0\} = F(t,p) \end{array} \qquad \text{(E2)}$$

where the notation $Y^{q-v} \equiv Y(t+(q-v)T,p)$ has been used to clarify the expression. Note that the Nth derivative has been grouped at the highest time steps. The coefficients b_v must be such that:

$$\sum_{v=0}^{q} b_v = 0 \quad , \sum_{v=0}^{q} \frac{Vb_2}{(N-1)} = 0 \quad , \; , \; , \sum_{v=0} \frac{(V)^{N-1} b_v}{(N-1)!} = 1 \qquad \text{(E3)}$$

in order to be consistent to $\partial^{N-1} Y/\partial t^{N-1}$ and q - N + 1 degrees of freedom are available to choose the b_v coefficients. The other approximations to $\partial^1 Y/\partial t^1$ must satisfy restrictions similar to eq. (E3) and q - i degrees of freedom are available to adjust the coefficients. The coefficients may be chosen to insure a stable difference scheme.

If eq. (E2) is multiplied by T^N, it is seen that the characteristic equation is:

$$\begin{array}{l} P(Z) = Z^q\{1 - \alpha_{N-1} T b_0 - \alpha_{N-2} T^2 C_0 + + + + +\} \\ \qquad +Z^{q-1}\{-N - \alpha_{N-1} T b_1 - \alpha_{N-2} T^2 C_1 + + + + +\} \\ \qquad \vdots \\ \qquad +Z^{q-N}\{(-1)^N - \alpha_{N-1} T b_N + + + + +\} \\ \qquad +Z^{q-N-1}\{-\alpha_{N-1} T b_{N+1} + + + + +\} \\ \qquad \vdots \end{array}$$

$$+Z^0\{-\alpha_{N-1}T\,b_q - \alpha_{N-2}T^2C_q + + + + - \alpha_0 T^N K_q\} \qquad \text{(E4)}$$

The roots of this complex polynomial must satisfy Von Neumann's Theorem in order for the difference equation to be stable. The known technique for testing the roots of P(Z) with respect to the unit circle of a complex plane is the Schur-Cohn criterion/11/17.

A simplification in examining the stability of eq. (E3) is that the expansion of the Schur-Cohn determinants (E4) yields real terms which can be arranged in ascending powers of T, T^2, T^3,.... This result is given as:

Proposition I. If the following conditions are met for the difference approximation to equation (A1):

(a) definite difference schemes are used to approximate the space derivatives

(b) the lowest power T-term of each Schur-Cohn determinant of the characteristic polynomial is of the proper algebraic sign and is nonzero except at frequences

$$p = 0,\ 2\pi/\Delta x,\ 4\pi/\Delta x,\ ,\ ,$$

then the difference approximation satisfies the Von Neumann necessary condition for stability as T→0 for a fixed Δx.

To see the basis for the proposition, note the use of definite difference schemes forces the lowest power T term in the Schur-Cohn determinants to approach zero from one direction only as p varies, $-\infty < p < \infty$. The odd derivatives contribute $\sin^{2s}(p\Delta x)$ as a factor, s = 0,1,2,..., because each determinant is real. With these two consequences of definite difference schemes, T can always be made small enough such that the lowest power of T dominates the determinant.

F - Properties of Time Polynomial Approximations

In this section the "invariants" of polynomial difference approximations are presented. The "invariants" are relationships which are true for all consistent approximations to an equation since they are limiting quantities which hold as T→0.

Consider first the three-step time difference approximation to a second order equation: (Note the signs of the α_i)

$$\begin{aligned}\frac{\partial^2 U}{\partial t^2} - \alpha_0 U \cong &\left\{\frac{1}{T^2} + \frac{b_1\alpha_1}{T} + C,\alpha_0\right\} U^{m+1}\\ &+\left\{\frac{-2}{T^2} + \frac{b_2\alpha_1}{T} + C_2\alpha_0\right\} U^{m}\\ &+\left\{\frac{1}{T^2} + \frac{b_3\alpha_1}{T} + C_3\alpha_0\right\} U^{m-1}\end{aligned} \qquad \text{(F1)}$$

Constraints

$$b_1+b_2+b_3=0$$

$$2b_1+b_2 =-1$$

$$C_1+C_2+C_3=-1$$

where the space dependency has been approximated and Fourier transformed prior to this step. The Schur-Cohn test yields "invariant" quantities which are independent of the free b_i and c_i:

$$\Delta_1 \cong T(\alpha_1+\bar{\alpha}_1),\quad \Delta_2 \cong T^4\{(\alpha_0-\bar{\alpha}_0)^2 - (\alpha_1+\bar{\alpha}_1)(\alpha_0\bar{\alpha}_1 + \bar{\alpha}_0\alpha_1)\} \qquad \text{(F2,F3)}$$

Proposition II. Every q + 1 time-step approximation to an Nth order equation in time has identically the same invariants for the Δ_{q-N+1}, Δ_{q-N+2}, Δ_{q-N+3}, ..., Δ_q Schur-Cohn determinants. Only the higher order T terms and the Δ_1, Δ_2, Δ_3, Δ_4, ..., Δ_{q-N} determinants vary with the type of approximation where $q \geq N$.

Since the "invariants" are fundamental to difference approximations, it is possible to determine apriori if a stable formulation can be found for an equation or system of equations by means of examining the "invariants." Such a classification is presented in Table I /12/ for the second order equation. In this table, there are eighty categories which depend on combinations of the real and imaginary parts of α_0 and α_1. These categories are divided into four groups. These groups are:

Table I CLASSIFICATION OF A SECOND ORDER SYSTEM

α_0 \ α_1	$Re\,\alpha_1=0$ $Im\,\alpha_1=0$	$Re\,\alpha_1\leq 0$ $Im\,\alpha_1=0$	$Re\,\alpha_1\leq 0$ $Im\,\alpha_1\gtrless 0$	$Re\,\alpha_1\leq 0$ $Im\,\alpha_1\lessgtr 0$	$Re\,\alpha_1=0$ $Im\,\alpha_1\gtrless 0$	$Re\,\alpha_1=0$ $Im\,\alpha_1\lessgtr 0$	$Re\,\alpha_1\geq 0$ $Im\,\alpha_1=0$	$Re\,\alpha_1\geq 0$ $Im\,\alpha_1\gtrless 0$	$Re\,\alpha_1\geq 0$ $Im\,\alpha_1\lessgtr 0$
$Re\,\alpha_0=0$ $Im\,\alpha_0=0$	╳	1	2	3	4	5	6 //////	7 ////// POSSIBLE	8 //////
$Re\,\alpha_0\leq 0$ $Im\,\alpha_0=0$	9	10 F.N.F.	11 F.N.F.	12 F.N.F.	13	14	15 //////	16 ////// FINAL	17 //////
$Re\,\alpha_0\leq 0$ $Im\,\alpha_0\gtrless 0$	18 =	19 F.N.F.	20 F.N.F.	21	22 =	23 =	24 //////	25 ////// VALUE	26 //////
$Re\,\alpha_0\leq 0$ $Im\,\alpha_0\lessgtr 0$	27 =	28 F.N.F.	29	30 F.N.F.	31 =	32 =	33 //////	34 ////// PROBLEMS	35 //////
$Re\,\alpha_0=0$ $Im\,\alpha_0\gtrless 0$	36 =	37 =	38 F.N.F.	39 =	40 =	41 =	42 =	43 =	44
$Re\,\alpha_0=0$ $Im\,\alpha_0\lessgtr 0$	45 =	46 =	47 =	48 F.N.F.	49 =	50 =	51 =	52	53 =
$Re\,\alpha_0\geq 0$ $Im\,\alpha_0=0$	54 //////	55 =	56 =	57 =	58	59	60 =	61 =	62 =
$Re\,\alpha_0\geq 0$ $Im\,\alpha_0\gtrless 0$	63 =	64 =	65	66 =	67 =	68 =	69 =	70 =	71 //////
$Re\,\alpha_0\geq 0$ $Im\,\alpha_0\lessgtr 0$	72 =	73 =	74 =	75	76 =	77 =	78 =	79 //////	80 =

Notes: [Highest even order of $\alpha_1]^2 \geq$ [odd order α_0];=(unstable); Im ( )$\lessgtr 0$ means f(pΔx)Sin(pΔx), with f(pΔx) ≥ 0

1) F.N.F. which means that any difference expression to eq. (F1) that results in the α_0 and α_1 listed is a stable formulation when T is sufficiently small that the invariant quantities dominate the Schur-Cohn determinants.

2) ——— means that the invariants are not the proper sign so that the difference equation is unstable for small T. Usually one root will be inside the unit circle and one root outside the unit circle.

3) (blank) means that combination of b_i and c_i may be chosen such that a formulation is found. The wave equation $\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$ is represented by category #9. Many of the implicit or explicit schemes used for this equation can be considered cases of particular c_i of eq. (F1).

4) ////// means that eq. (F2) is greater than zero, and both roots of the characteristic equation lie outside the unit circle. In this circumstance, it may be possible to formulate the difference equation as a final-value problem and proceed to solve the equation backwards in time. Laplace's equation, $\partial^2 u/\partial t^2 = -\partial^2 u/\partial x^2$, is one example which may be represented by category #54.

It is possible to extend to higher order equations the classification of approximation as done in Table I, but the number of terms in the invariants increase with the order of the equation and such a classification becomes unweildy above the third order /13/. In the special case where the α_i' are real (the equation contains only even order space derivatives), it may be easily verified that the invariants reduce to the stability conditions given by Jury's Stability Table /8/ for sampled data systems.

G - Conclusions and Discussion

The development of discrete forms for systems described by partial differential equations has shown the point where the separation begins from the analysis of ordinary differential systems, mainly in approximating the operator $\exp(\lambda(p)T)$. The variety of difference equations for a given partial differential equation represent attempts to approximate the operator.

The general approach of this paper is a method which yields many of the particular approximations to a partial differential equations plus is an a-priori criterion to determine if a stable difference scheme may be formulated. The method has been applied to a second order optimal control problem /14/ and boundary control of distributed parameter systems /15/, the latter example in excellent agreement with analytical results obtained by other methods.

REFERENCES

1. C.J. Tranter, "Integral Transforms in Mathematical Physics," John Wiley, New York, 1951.
2. Birkhoff, Garrett,"Classification of Partial Differential Equations," Journal SIAM. Vol. 2, No. 1, March, 1954.
3. F. L. Alt, "Electronic Digital Computers," Section 4.7 on partial differential equations, Academic Press, New York, 1958.
4. R. Courant, "Hyperbolic Partial Differential Equations and Applications," in Modern Mathematics for the Engineers edited by E.F.Beckenbach, Univ. of Cal. Engineering, Extension Series, McGraw-Hill,New York, 1956.
5. Kirstein P.T.,"On the Determination of the Electrodes Required to Produce a Given Electric Field Distribution Along a Prescribed Curve," Proceedings of the IRE, Oct. 1958.
6. Harker, K.J.,"Determination of Electrode Shapes for Axically Synthetric Electron Guns,"Journal of Applied Physics, Dec. 1960, Page 2165.
7. W.R. Perkins and I.B.Cruz,"Engineering of Dynamic Systems,"John Wiley, New York, 1969.
8. E. I. Jury, "Theory and Application of the Z-transform," John Wiley, New York, 1964.
9. R.D. Richtmyer,"Difference Methods for Initial-Value Problems," Interscience Publishers, New York, 1957, or the second edition by R.D. Richtmyer and K. W. Morton, Interscience Publishers, New York, 1967.
10. M.G. Salvadori and M.L. Baron, "Numerical Methods in Engineering," Prentice-Hall, Englewood Cliffs, New Jersey, 1952.
11. M.Marden, "The Geometry of the Zeros," American Mathematical Society, New York, 1949.
12. G.L. Kusic, "On Stability of Numerical Methods for Systems of Initial-Value Partial Differential Equations,"Journal of Computational Physics, Vol. 4, No. 2, August 1969.
13. G.L.Kusic and K.Cooper, "Stability of Difference Approximations for Third Order Initial-Value Partial Differential Equations,"to be published.
14. G.L.Kusic,"Finite Differences to Implement the Solution for Optimal Control of Distributed Parameter Systems," IEEE Trans. Automatic Control, Vol. AC-15, August 1969.
15. B. Lazarevic and G.L. Kusic, "Optimal Boundary Control of Distributed Parameter Systems,"Boris Kidric Institute of Nuclear Science, Report IBK-1006, Beograd, Yugoslavia, March 1971.
16. Forsythe and Wasov, "Finite Difference Methods for Partial Differential Equations," J. Wiley, New York, 1960.
17. Kusic, G.L.,"Implementing the Solution to Continuous Optimal Control of Distributed Parameter Systems by Finite Differences." Proceedings of the Sixth Annual Allerton Conference on Circuit and System Theory,October, 1968.

A SENSOR LOCATION ALGORITHM FOR DISTRIBUTED-PARAMETER SYSTEMS

DONALD J. EWING, JR.
University of Toledo
Toledo, Ohio 43606

THOMAS J. HIGGINS
University of Wisconsin
Madison, Wisconsin 53706

ABSTRACT

This paper demonstrates the use of an algorithm for optimal sensor placement in distributed-parameter systems. The algorithm is based on approximation of the system state using variational principles for the solution of partial differential equations. Examples illustrate application of the method to systems described by the wave equation and the diffusion equation.

INTRODUCTION

Optimal control strategies for distributed-parameter systems require, in general, data on system state at every point in a system. Since this is obviously impossible in any physical realization, careful attention must be given to placement of sensors. Some important aspects of the realization of distributed-parameter control systems have been outlined by Athans [1]. Among the problems outlined by Athans are:

(a) methods which allow sensor numbers and locations to be left as unspecified parameters until the optimization calculations are completed, and
(b) optimization of the number and location of measuring instruments.

Kaiser [2] presents an assessment of the current state of heater-sensor design techniques for thermal control problems.

The work summarized here applies variational techniques to solution of the dynamic system equations. Applications of variational techniques to the solution of differential equations is a long established method [3,4]. Noble [5] has demonstrated the utility of variational methods for approximate solution of differential equations using arbitrarily chosen trial functions.

Although many works on optimization and state estimation employ system eigenfunctions, this can be quite awkward because of the computational time required by manipulations involving functions such as trigonometric and Bessel functions. The burden is especially heavy for small control computers with limited hardware. Ordinary polynomials, for instance, involve considerably less computing time when functions are fitted to data at a set of points.

The problem which evolves, then, in sensor placement is how to best fit system dynamic responses with arbitrary functions. The solution offered here is approximate solution of system dynamic equations using the calculus of variations. Based upon "exact" calculation, the process used proceeds as follows:

(a) choose a set of sensor locations,
(b) compute the set of sensor outputs as a function of time,
(c) fit a function of suitable form to the data, and
(d) evaluate the functional for the variational principle used.

This procedure is embedded in a parameter optimization algorithm which treats sensor locations as search parameters.

RECASTING A PROBLEM IN PARTIAL DIFFERENTIAL EQUATIONS AS A PROBLEM IN THE CALCULUS OF VARIATIONS

This section outlines methods for deriving variational principles. The methods are those of Noble [5,6]. Noble's procedures are most helpful, since they give a systematic procedure for approaching a problem which is largely treated as a black art. A more general treatment of the problem is given in Rall [7]. Komkov [8] applies the methods of Noble and Rall to a problem in elastodynamics.

Noble's procedure begins with the functional

$$J = \langle Lc, p\rangle - W(c,p). \tag{1}$$

Equation (1) may also be written in the equivalent form shown in equation (2). The stationary solution which extremizes the functionals

$$J = \langle c, L^*p\rangle - W(c,p) \tag{2}$$

is given in

$$Lc = \partial W/\partial p \tag{3}$$

and

$$L^*p = \partial W/\partial c. \tag{4}$$

Assuming that c is the problem variable of interest, L and W are chosen to give the variable p as a function of c using equation (3). Substitution of p in equation (4) yields the equation which, if the choices have been made properly, is that whose solution is sought. Since simultaneous solution of equations (3) and (4) are the conditions for an extremum of J, the extremizing function is a solution of the equation of interest. Although the formal adjoint is all that has received attention in this discussion, attention must be given to boundary requirements of the optimal solution. Should the boundary terms required not be compatible, it is frequently possible to add boundary terms to J which will result in acceptable solutions.

For the particular case in which $W(c,p)$ is a quadratic functional defined by the scalar product $\langle,\rangle$ more specific results may be written. In this case the partial derivatives of equations (3) and (4) may be written as in

$$\partial W/\partial c = f_1 + Mc + Np \tag{5}$$

and

$$\partial W/\partial p = f_2 + Yc + Zp \tag{6}$$

where the quantities on the right hand sides designated by capital letters are transformation operators. Completing the equalities of equations (3) and (4) leads to

$$f_1 + Mc + Np = L^*p \tag{7}$$

and

$$f_2 + Yc + Zp = Lc. \tag{8}$$

Establishing a relationship between the costate variable, p, and c from equation (8) gives

$$Zp = Lc - Yc - f_2. \tag{9}$$

The transforms L* and N are redefined as in

$$L^* = XZ \tag{10}$$

and

$$N = VZ \tag{11}$$

to avoid use of the inverse Z operator. From the redefinitions of equations (10) and (11), and substituting equation (9) in equation (7):

$$(V - X)(L - Y)c + Mc + (f_1 - Vf_2 - Xf_2) = 0. \tag{12}$$

This equation represents the differential equation which is to be solved by variational methods. As an illustration of this procedure, consider the following examples.

A Variational Principle for the Wave Equation

The example to be considered here is formulation of a variational principle for solution of the wave equation,

$$\partial^2 c/\partial t^2 = \nabla^2 c. \tag{13}$$

Even with the assistance of the equations here, it is apparent that substantial intuition is involved. Initial guesses for several of the operators as $L \equiv \partial/\partial t$, $M \equiv -\nabla^2$ and $Z = 1$ result in

$$(V \frac{\partial}{\partial t} - VY)c - \frac{\partial}{\partial t} Yc + \frac{\partial^2 c}{\partial t^2} - \nabla^2 c = Vf_2 + Xf_2 - f_1. \tag{14}$$

By this point it is clear that the wave equation results for $V=Y=f_2=f_1=0$. With these substitutions, equation (14) becomes equation (13).

For the above choice of operators, equations (5) and (6) become

$$\partial W/\partial c = -\nabla^2 c \tag{15}$$

and

$$\partial W/\partial p = p. \tag{16}$$

By use of the anti-derivative of the Frechet derivatives, and the fact that the adjoint of the operator $\nabla\cdot$ is $-\nabla$, equations (15) and (16) show that W(c,p) is given by

$$W(c,p) = \frac{1}{2} <\nabla c,\nabla c> + \frac{1}{2} <p,p>. \tag{17}$$

The complete functional, J, is

$$J = <\frac{\partial c}{\partial t},p> - \frac{1}{2} <\nabla c,\nabla c> - \frac{1}{2} <p,p>. \tag{18}$$

The costate variable, p, in this problem is the time derivative of c.

Before the complete solution of a specific problem can be obtained, an investigation of boundary conditions for existence of an optimal solution is necessary. Certain surface integrals may need to be added to equation (18) to handle these problems.

The scalar product in this case is given by

$$<x,y> \;=\; \int_0^T \int_V xy \; dv\; dt. \tag{19}$$

A Variational Principle for the Diffusion Equation

System dynamics are, for this case, given by

$$\partial c/\partial t \;=\; \nabla^2 c + f. \tag{20}$$

Equation (20) may be written in terms of the variables p and c, as in

$$p \;=\; \nabla^2 c + f \tag{21}$$

and

$$c^{(1)} \;=\; p. \tag{22}$$

Applying the methods above, a guess is made that $L = \partial/\partial t$. From this, equations (3) and (4) may be written as

$$c^{(1)} \;=\; \partial W/\partial p \tag{23}$$

and

$$p^{(1)} \;=\; -\partial W/\partial c. \tag{24}$$

Examination of equations (21) and (22) shows that they may be written as

$$p^{(1)} \;=\; \nabla^2\, c^{(1)} + f^{(1)} \tag{25}$$

and

$$c^{(1)} \;=\; p. \tag{26}$$

From equation (20), however, it is seen that equation (25) may be written as

$$p^{(1)} \;=\; \nabla^2[\nabla^2\, c + f] + f^{(1)}. \tag{27}$$

Substitution from equations (26) and (27) in equations (23) and (24) result in derivatives of $W(c,p)$ which are

$$p \;=\; \partial W/\partial p \tag{28}$$

and

$$-\nabla^2\nabla^2 c - \nabla^2 f - f^{(1)} \;=\; \partial W/\partial c. \tag{29}$$

Integration of equations (28) and (29) result in

$$W(c,p) = \tfrac{1}{2}<p,p> - \tfrac{1}{2}<\nabla^2 c,\nabla^2 c> - <\nabla^2 c,f> - <c,f^{(1)}>. \tag{30}$$

Using equation (30) in equation (1) gives the variational expression

$$J = <p,c^{(1)}> - \tfrac{1}{2}<p,p> + \tfrac{1}{2}<\nabla^2 c,\nabla^2 c> + <\nabla^2 c,f> + <c,f^{(1)}>. \tag{31}$$

Manipulation of the last term in equation (31) gives, as an equivalent form,

$$J = <p,c^{(1)}> - \tfrac{1}{2}<p,p> + \tfrac{1}{2}<\nabla^2 c,\nabla^2 c> + <\nabla^2 c-c^{(1)},f>. \tag{32}$$

Choice of a trial variable p satisfying equation (3) gives, when substituted in equation (32), a variational expression

$$J_1 = \tfrac{1}{2}<c^{(1)},c^{(1)}> + \tfrac{1}{2}<\nabla^2 c,\nabla^2 c> + <\nabla^2 c-c^{(1)},f>. \tag{33}$$

Before application of this variational formulation can be carried out, a thorough examination of the boundary conditions required is necessary. To accomplish this examination, it is necessary to write the scalar products in their integral form. Expanding

$$J_1 = \int_0^T \int_V [\tfrac{1}{2}(c^{(1)})^2 + \tfrac{1}{2}(\nabla^2 c)^2 + f\nabla^2 c - c^{(1)} f]\ dv\ dt \tag{34}$$

about the optimal c, where c is given by

$$c = c_o + \theta\eta \tag{35}$$

leads to

$$\begin{aligned} J_1 = &\int_0^T \int_V [\tfrac{1}{2}\, c_o{}^{(1)2} + \tfrac{1}{2}(\nabla^2 c_o)^2 + f\ \nabla^2\ c_o - c_o{}^{(1)} f]\ dv\ dt \\ &+ \theta \int_0^T \int_V [c_o{}^{(1)}\eta^{(1)} + \nabla^2 c_o\ \nabla^2\eta + f\nabla^2\eta - \eta^{(1)} f]\ dv\ dt \\ &+ \theta^2 \int_0^T \int_V [\tfrac{1}{2}\eta^{(1)2} + \tfrac{1}{2}(\nabla^2\eta)^2]\ dv\ dt. \end{aligned} \tag{36}$$

The necessary condition for a stationary solution is that the second term in equation (36) be zero, as shown in

$$\int_0^T \int_V [c_o{}^{(1)}\eta^{(1)} + \nabla^2 c_o\ \ \nabla^2\eta + f\nabla^2\eta - \eta^{(1)} f]\ dv\ dt = 0. \tag{37}$$

Taking the adjoint of operators acting on the variational term, η, gives

$$\begin{aligned} &\int_0^T \int_V \eta[-c_o{}^{(2)} + \nabla^2\nabla^2 c_o + \nabla^2 f + f^{(1)}]\ dv\ dt + \int_V \eta c_o{}^{(1)}\Big|_0^T dv \\ &+ \int_0^T \int_S n\cdot(\nabla^2 c_o\ \nabla\eta - \eta\nabla\ \nabla^2 c_o)\ ds\ dt + \int_0^T \int_S n\cdot(f\nabla\eta - \eta\nabla f)\ ds\ dt \\ &- \int_V \eta f\Big|_0^T dv = 0. \end{aligned} \tag{38}$$

Conditions for a stationary solution over the space of the system are given by setting the first term equal to zero. Testing the bracketted term for conditions under which it is zero show the equality to be satisfied for

$$c_o{}^{(1)} = \nabla^2 c_o + f. \tag{39}$$

Both the second and last terms in equation (38) present potential difficulties in achieving stationary conditions. Addition of two terms to the functional, J_1, removes the potential difficulties of the above named terms. The resulting functional is shown in equation (40). Although the added terms require knowledge of the optimal solution at

$$\Phi(c) = J_1(c) - \int_V c\ c_o{}^{(1)}\Big|_0^T dv + \int_V cf\Big|_0^T dv \tag{40}$$

$t = T$ and $t = 0$, such knowledge is available in the intended application. The only requirement on c must be that natural boundary conditions are necessary over the spatial region of the system.

APPLICATION OF VARIATIONAL METHODS TO SENSOR PLACEMENT

Sensors may be placed in optimal locations for a system by choosing a set of locations giving a minimum of the functional in the variational formulation. For the system described by the diffusion equation, as in the example above, the functional is

$$\Phi(c) = \int_0^T \int_V [\tfrac{1}{2}(c^{(1)})^2 + \tfrac{1}{2}(\nabla^2 c)^2 + f\nabla^2 c - c^{(1)} f]\, dv\, dt$$
$$- \int_V c\, c_0^{(1)} \Big|_0^T dv + \int_V cf \Big|_0^T dv. \qquad (41)$$

The trial function, c(x,t), is computed using a function, chosen for convenience, computed from sensor outputs. One such function might be a polynomial of the form

$$c(x,t) \approx \sum_{i=0}^{n} a_i(t)\, x^i \qquad (42)$$

where the coefficients, a_i, are computed from the outputs at each sampling time. As a more general expression, the function c is taken as approximated by

$$c(x,t) \approx \sum_{i=0}^{n} \alpha_i(t)\, g_i(x). \qquad (43)$$

Equation (41) requires that a trial function, c, be once differentiable in time. The derivative terms may, therefore, be approximated by assuming a piecewise linear function of time. Similar requirements are seen for the case of the wave equation, equation (18). This approximation is especially suited to computer-controlled systems with data sampling, where derivatives are computed using successive sample values. Deviation of the functional from the minimal is a measure of combined approximation errors in the spatial and time coordinates.

The nature of the approximating functions, g, order of the approximations (number of sensors) and sampling intervals may be varied until a satisfactorily accurate approximation is achieved.

An illustration of the methods is given in the following example.

Sensor Location in a Diffusion System

As an illustration, consider the problem of optimally placing sensors to approximate the state as a function of time and displacement when the system is characterized by

$$\partial c/\partial t = \partial^2 c/\partial x^2 \; ; \qquad 0 \le x \le \pi. \qquad (44)$$

The initial conditions are assumed to be

$$c(0,x) = \sum_i a_i \sin ix \; ; \quad i \text{ an odd integer}. \qquad (45)$$

An observation period of 1-second is used, with samples at $t = 0$ and intervals of h seconds thereafter. Boundary conditions are $c(t,0) = 0$ and $c(t,\pi) = 0$.

Approximation of the spatial distribution is by the polynomial

$$c(kh,x) = \sum_{i=0}^{n} A_i(kh)\, x^i. \tag{46}$$

The variational expression for this problem is, from equation (41),

$$\Phi(c) = \int_0^1 \int_0^\pi \left[\frac{1}{2}\left(\frac{\partial c}{\partial t}\right)^2 + \frac{1}{2}\left(\frac{\partial^2 c}{\partial x^2}\right)^2\right] dx\, dt - \int_0^\pi c\, \frac{\partial c_o}{\partial t}\Big|_0^T dx. \tag{47}$$

Placement of sensors for this problem is accomplished by a steepest descent algorithm which incorporates equation (47) as a cost functional. The search vector is the vector of sensor locations. A recursion formula of the form in equation (48) forms the basic algorithm,

$$x_s(j+1) = x_s(j) + m\, \partial\Phi/\partial x_s \tag{48}$$

where $x_s(j)$ is the sensor location vector at the j-th iteration.

The steepest descent program calls a subroutine which generates the cost functional each time it is required. The main program delivers sensor locations and initial state data to the subroutine. A flow-chart, Figure 1, shows the steps involved in computation of the functional.

Some constraints on trial solutions are imposed by the boundary requirements of equation (38). From the third term in equation (38) it is seen that $c(t,x)$ is subject to essential boundary conditions at $x = 0$ and $x = \pi$. This implies, therefore, that sensors are required at $x = 0$ and $x = \pi$. As $\nabla^2 c_o$ is zero at the boundaries, the boundary conditions on the gradient are natural conditions. No constraints are imposed on this function. Differentiability requirements demand that any trial function be twice differentiable in the spatial variables. The latter requirement demands at least second order polynomials, which implies at least 3 sensors.

Optimal sensor locations are presented for two sets of initial states. The first numerical calculation is for the initial-state

$$c(0,x) = 3 \sin x. \tag{49}$$

A sampling interval of 0.1-second is used. The numerical value of the minimal functional, Φ_o, is -6.20. A tabulation of numerical results is given in Table 1. For the case of 5 sensors it is seen that the minimum functional is -6.098, which is approaching the true minimum.

The second numerical illustration is for the initial state

$$c(0,x) = 3 \sin x + 1 \sin 3x. \tag{50}$$

A sampling interval of 0.1-second is used here, too. A tabulation of results for this case is given in Table 2.

SUMMARY

This paper presents a technique for optimal location of sensors for a specified state estimation strategy for distributed-parameter systems. Quantitative comparisons are readily made between alternate estimation schemes, and between different numbers of sensors. At the heart of the technique is a variational method which furnishes the quantitative measure of the "goodness" of estimation. The method is independent of whether control is distributed or applied at boundaries, although details in the analysis are affected.

REFERENCES

1. M. Athans, "Some Remarks on the Control of Distributed-Parameter Systems," PROCEEDINGS, 1969 JACC, University of Colorado, Boulder, Colorado, August, 1969, pp. 615-617.

2. K. W. Kaiser, "A Method of Determining the Heater-Sensor Configuration in Temperature Distribution Controllers," PROCEEDINGS, IFAC SYMPOSIUM ON THE CONTROL OF DISTRIBUTED PARAMETER SYSTEMS, Banff, Alberta, Canada, June, 1971, paper 1-3.

3. F. B. Hildebrand, Methods of Applied Mathematics, Prentice-Hall, Inc., New York, 1952.

4. R. Courant, D. Hilbert, Methods of Mathematical Physics - Vol. I, Interscience Publishers, Inc., New York, 1953.

5. B. Noble, "Complementary Variational Principle for Boundary Value Problems I. Basic Principles with an Application to Ordinary Differential Equations," TECHNICAL SUMMARY REPORT NO. 473, Army Mathematics Research Center, University of Wisconsin, Madison, Wisconsin, November, 1964.

6. B. Noble, Applied Variational Methods (Unpublished Notes), Department of Mathematics, University of Wisconsin, Madison, Wisconsin, January, 1963.

7. L. B. Rall, "On Complementary Variational Principles," JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATION, Vol. 14, No. 2, May, 1966, pp. 174-184.

8. V. Komkov, "Application of Rall's Theorem to Classical Elastodynamics," JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATION, Vol. 14, No. 3, June, 1966, pp. 511-521.

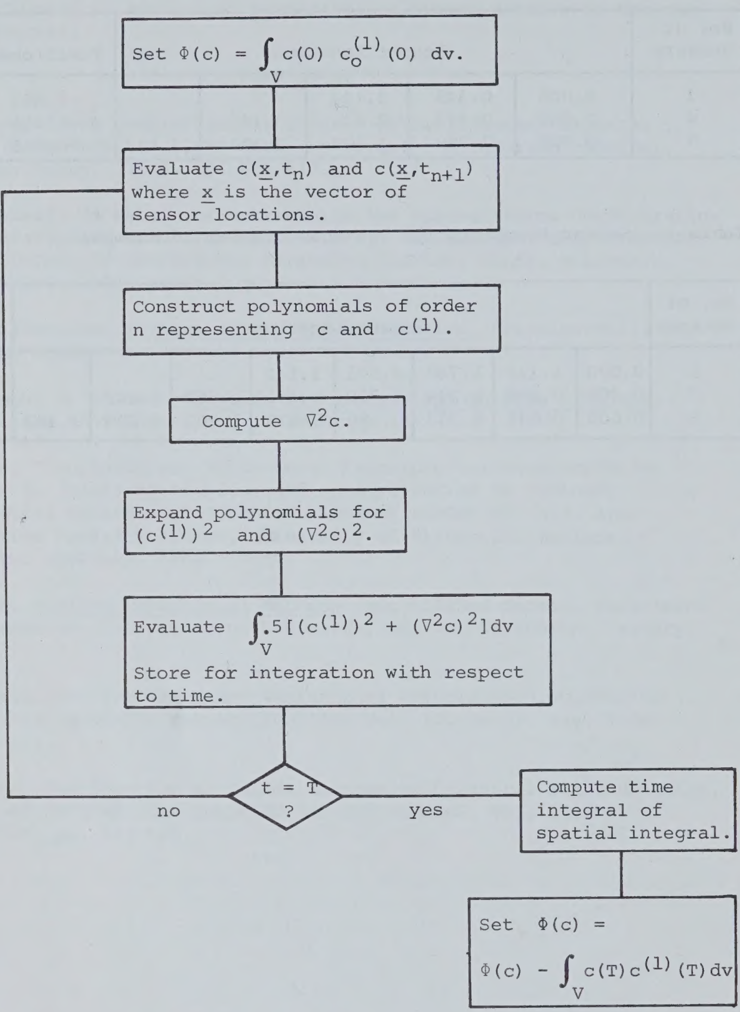
Table 1: Sensor Locations for Single Mode Problem of Example

No. of Sensors	Sensor Locations					Functional
3	0.000	0.349	3.142			-5.461
4	0.000	0.472	2.640	3.142		-5.455
5	0.000	0.750	1.500	2.300	3.142	-6.098

Table 2: Sensor Locations for Two Mode Problem of Example

No. of Sensors	Sensor Locations								Functional
5	0.000	1.117	1.780	2.521	3.142				-25.7
7	0.000	0.898	1.204	1.710	2.152	2.562	3.142		-147.5
8	0.000	0.045	0.333	0.596	0.819	1.802	2.298	3.142	-172.7

Figure 1: Flowchart of Subprogram for Functional Evaluation for Example


Set $\Phi(c) = \int_V c(0)\, c_o^{(1)}(0)\, dv$.
Evaluate $c(\underline{x},t_n)$ and $c(\underline{x},t_{n+1})$ where $\underline{x}$ is the vector of sensor locations.
Construct polynomials of order n representing c and $c^{(1)}$.
Compute $\nabla^2 c$.
Expand polynomials for $(c^{(1)})^2$ and $(\nabla^2 c)^2$.
Evaluate $\int_V .5[(c^{(1)})^2 + (\nabla^2 c)^2]dv$
Store for integration with respect to time.
t = T ?
no
yes
Compute time integral of spatial integral.
Set $\Phi(c) =$
$\Phi(c) - \int_V c(T)c^{(1)}(T)dv$

BOUNDARY FILTER FOR LINEAR DISTRIBUTED PARAMETER SYSTEMS

S.R. ATRE AND S.S. LAMBA
Department of Electrical Engineering
Indian Institute of Technology
New Delhi-29, India

ABSTRACT

The problem of estimating the state of linear distributed parameter systems from boundary measurements is treated. A direct development of filtering and variance equations are presented via innovations theory. They represent the boundary version of the distributed Kalman-Bucy filter.

INTRODUCTION

The study of distributed state estimation problem has been receiving considerable attention in recent years. The purpose of the estimator for distributed parameter systems is to continually update the partial differential equation model of the system based on noisy output measurements. This in turn leads to controller action based on most currently realised model.

The dynamic behaviour of the distributed parameter system distinguishes the boundary region from its domain. Earlier developments on filtering for distributed parameter systems concentrated mostly on measurement data obtained from the interior of the domain. Domain measurements considered in [1] and [2] are obtained by fixing the probes over the spatial locations in the entire interior. Thau [3] uses a single point measuring device placed at a convenient point inside the domain. However, in the case of a single probe, the data scanning over the whole domain is suggested [4].

To be more realistic the focus should be on boundary measurements rather than over its domain [5]. With system accessibility restricted to boundary the control action has to be achieved via boundary controllers. In such situations the possibility of system disturbances entering through the boundary cannot be ruled out. Previous study pertaining to boundary measurements has been carried out by Balkrishnan and Lions [6]. However their development was restricted to noise free boundary conditions.

In this paper the distributed parameter model in terms of partial differential equation is considered. The disturbance at the boundary appear as nonhomogeneous boundary conditions in the mathematical model of the system. Measurement data obtained from boundary are assumed noisy and are treated as an additive vector in the measurement equation. Application of the innovations approach [7], leads to a Kalman-Bucy type of filter. This filter acts directly on the boundary data and is therefore termed as 'boundary filter'.

PROBLEM FORMULATION

Let Ω be the domain of an n-dimensional Euclidean space E^N with boundary $\partial\Omega$. Consider the linear distributed system described by

$$\frac{\partial X(x,t)}{\partial t} = S_x X(x,t) \qquad x \varepsilon \Omega \qquad (1)$$

where $X(x,t)$ is the n-dimensional state function of the system and S_x is an $n \times n$ matrix, linear spatial operator.

The boundary conditions for eqn. (1) can, in general be homogeneous. By considering a stochastic term, we account for the internal system disturbances entering through boundary. As such the boundary conditions become inhomogeneous, viz.

$$L_\beta X(\beta,t) = G(\beta,t) W(\beta,t) \qquad \beta \varepsilon \partial \Omega \qquad (2)$$

where L_β is the spatial operator on boundary, $G(\beta,t)$ is the known input matrix and $W(\beta,t)$ is a zero-mean white noise with covariance

$$E\{W(\beta,t) W'(\alpha,\tau)\} = Q(\beta,\alpha,t)\, \delta(t-\tau) \qquad \beta,\alpha \varepsilon \partial \Omega \qquad (3)$$

where E{ } indicates the expectation operator and "′" denotes transpose.

The measurement process on boundary can be described by the following relation:

$$Z(\beta,t) = H(\beta,t) X(\beta,t) + V(\beta,t) \qquad \beta \varepsilon \partial \Omega \qquad (4)$$

where $Z(\beta,t)$ is an m-vector, $H(\beta,t)$ is an $m \times n$ output matrix and $V(\beta,t)$ is the zero-mean white noise with covariance

$$E\{V(\beta,t) V'(\alpha,\tau)\} = R(\beta,\alpha,t)\, \delta(t-\tau) \qquad \beta,\alpha \varepsilon \partial \Omega \qquad (5)$$

The boundary filtering problem consists in obtaining an optimal estimate $\hat{X}(x,t/t)$ of $X(x,t)$ based on measured data $Z(\beta,\tau) : \beta \varepsilon \partial \Omega,\ t_o \leqslant \tau \leqslant t$.

If $\tilde{X}(x,t/t) = X(x,t) - \hat{X}(x,t/t)$, represents the estimation error, then the optimal estimate can be defined to satisfy the following relation:

$$E\{\tilde{X}(x,t/t)\, Z'(\alpha,\tau)\} = 0 \qquad (6)$$

LINEAR FILTERING

The relation (6) simply means that the estimation error is uncorrelated with measurement data obtained from the boundary. In the sequel eqn. (6) can also be put as

$$E\{\tilde{X}(x.t/t)\, \nu'(\alpha,\tau)\} = 0 \qquad (7)$$

where $\nu(..)$ is the innovation process [7] conveying same statistical information as $Z(..)$, viz.

$$\nu(\beta,t) = Z(\beta,t) - H(\beta,t)\,\hat{X}(\beta,t/t) \qquad \beta \in \partial\Omega \tag{8}$$

and $$E\{\nu(\beta,t)\ \nu'(\alpha,\tau)\} = R(\beta,\alpha,t)\ \delta(t-\tau) \tag{9}$$

By considering the optimal estimate to be linear operation on measured data and using the equivalence between $Z(..)$ and $\nu(..)$, $\hat{X}(..)$ can be expressed as

$$\hat{X}(x,\ t/t) = \int_{t_0}^{t}\int_{\partial\Omega} M(x,t;\eta,\tau)\nu(\eta,\tau)\ d(\partial\Omega_\eta)d\tau \tag{10}$$

Thus the desired estimate can be obtained once we know the Kernel matrix $M(..,..)$ Using eqns. (7) and (9), eqn (10) becomes

$$\hat{X}(x,t\,/\,t) = \int_{t_0}^{t}\int_{\partial\Omega}\int_{\partial\Omega} E\{X(x,t)\ \nu'(\eta,\tau)\}R^{i}(\eta,\alpha,\tau). \\ .\nu(\alpha,\tau)\ d(\partial\Omega_\eta)\ d(\partial\Omega_\alpha)d\tau \tag{11}$$

where the index 'i' over the matrix $R(...)$ denotes its inverse defined by the relation

$$\int_{\partial\Omega} R(\beta,\eta,t)\ R^{i}(\eta,\alpha,t)\ d(\partial\Omega_\eta) = \delta(\beta-\alpha)\ I \tag{12}$$

The eqn. (11) is the general formula for the linear estimate of $X(x,t)$ based on boundary measurement data.

KALMAN-BUCY STRUCTURE

In order to obtain the boundary version of the distributed Kalman-Bucy filter the eqn. (11) is partially differentiated with respect to t to yield

$$\frac{\partial\hat{X}(x,t/t)}{\partial t} = S_x\hat{X}(x,t/t) + \int_{\partial\Omega}\int_{\partial\Omega} E\{X(x,t)\ \nu'(\eta,t)\}R^{i}(\eta,\alpha,t). \\ .\nu(\alpha,t)\ d(\partial\Omega_\eta)\ d(\partial\Omega_\alpha) \tag{13}$$

In eqn. (13), the term corresponding to Kalman gain can be put as

$$\begin{aligned} K(x,t,\alpha,t) &= \int_{\partial\Omega} E\{X(x,t)\ \nu'(\eta,t)\}\ R^{i}(\eta,\alpha,t)\ d(\partial\Omega_\eta) \\ &= \int_{\partial\Omega} E\{X(x,t)[\tilde{X}'(\eta,t/t)H'(\eta,t)+V'(\eta,t)]\}R^{i}(\eta,\alpha,t)d\Omega_\eta \\ &= \int_{\partial\Omega} E\{[\hat{X}(x,t/t)+\tilde{X}(x,t/t)]\ \tilde{X}'(\eta,t/t)\}H'(\eta,t). \\ &\qquad\qquad \cdot R^{i}(\eta,\alpha,t)d(\partial\Omega_\eta) \\ &= \int_{\partial\Omega} P(x,\eta,t)H'(\eta\ ,t)R^{i}(\eta,\alpha,t)d(\partial\Omega_\eta) \end{aligned} \tag{14}$$

where P(...) denotes the error covariance given by

$$P(x,\eta,t) = E\{\tilde{X}(x,t/t)\ \tilde{X}'(\eta,t/t)\} \qquad (15)$$

Using (14) the distributed filter of eqn. (13) becomes

$$\frac{\partial\hat{X}(x,t/t)}{\partial t} = S_x\hat{X}(x,t/t)+\int_{\partial\Omega}\int_{\partial\Omega}P(x,\eta,t)H'(\eta,t)R^{i}(\eta,\alpha,t).$$

$$.\ v(\alpha,t)d(\partial\Omega_\eta)d(\partial\Omega_\alpha) \quad x\ \varepsilon\ \Omega \qquad (16)$$

From (2), the boundary condition for eqn. (16) canbe set as

$$L_\beta\ \hat{X}(\beta,t/t) = 0 \qquad \beta\ \varepsilon\ \partial\ \Omega \qquad (17)$$

As such the estimation error $\tilde{X}(x,t/t)$ satisfies the following partial differential equation

$$\frac{\partial\tilde{X}(x,t/t)}{\partial t} = S_x\ \tilde{X}(x,t/t) - \int_{\partial\Omega}\int_{\partial\Omega}P(x,\eta,t)H'(\eta,t)R^{i}(\eta,\alpha,t).$$

$$.H(\alpha,t)\tilde{X}(\alpha,t/t)d(\partial\Omega_\eta)d(\partial\Omega_\alpha)$$

$$-\int_{\partial\Omega}\int_{\partial\Omega}P(x,\eta,t)H(\eta,t)R^{i}(\eta,\alpha,t)V(\alpha,t)d(\partial\Omega_\eta)d(\partial\Omega_\alpha) \qquad (18)$$

with boundary condition

$$L_\beta\ \tilde{X}(\beta,t/t) = G(\beta,t)\ W(\beta,t) \qquad (19)$$

Using (18) and (19) with modest amount of effort, this leads to following Riccati partial differential equation for P(...)

$$\frac{\partial P(x,y,t)}{\partial t} = S_x\ P(x,y,t) + P(x,y,t)S_y'$$

$$-\int_{\partial\Omega}\int_{\partial\Omega}P(x,\eta,t)H'(\eta,t)R^{i}(\eta,\alpha t)H(\alpha,t)P(\alpha,y,t).$$

$$.d(\partial\Omega_\eta)\ d(\partial\Omega_\alpha)$$

$$+\int_{\partial\Omega}\int_{\partial\Omega}B(x,t;\eta,t)G(\eta,t)Q(\eta,\alpha,t)G'(\alpha,t).$$

$$.B'(y,t;\alpha,t)d(\partial\Omega_\eta)d(\partial\Omega_\alpha)$$

$$x,y,\varepsilon\ \Omega \qquad (20)$$

with boundary condition

$$L_\beta\ P(\beta,y,t) = \int_{\partial\Omega}G(\beta,t)Q(\beta,\alpha,t)G'(\alpha,t)B'(y,t;\alpha,t)d(\partial\Omega_\alpha)$$

$$\beta\ \varepsilon\ \partial\ \Omega,\ y\ \varepsilon\ \Omega \qquad (21)$$

where B(..,..) denotes the Green's kernel onthe boundary of the system described by eqns. (1) and (2).

The equations (16)-(17) and (20)-(21) represent the boundary filter with Kalman-Bucy structure. Fig. 1 is the block diagram representation of the distributed filter.

EXAMPLE

Consider the scalar parabolic system governed by the equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} \qquad 0 < x < 1 \tag{22}$$

subject to boundary conditions

$$\left.\frac{\partial u(x,t)}{\partial x}\right|_{\substack{x = o \\ x = 1}} = 0 \tag{23}$$

We assume that for estimating $u(x,t): x \varepsilon \Omega$, a probe is fixed at the origin of x axis (which is at the boundary) and the measurement process becomes

$$Z(o,t) = u(o,t) + v(t) \tag{24}$$

where $v(t)$ is assumed to be the zero-mean white noise with covariance

$$E\{v(t)\ v(\tau)\} = r\ \delta(t-\tau) \tag{25}$$

For the single point measurement at the boundary the filter and variance eqns. are given by

$$\frac{\partial \hat{u}(x,t/t)}{\partial t} = \frac{\partial^2 \hat{u}(x,t/t)}{\partial x^2} + P(x,o,t)(1/r)[Z(o,t)-\hat{u}(o,t/t)] \qquad x \varepsilon \Omega \tag{26}$$

$$\left.\frac{\partial \hat{u}(x,t/t)}{\partial x}\right|_{\substack{x = o \\ x = 1}} = 0 \tag{27}$$

and

$$\frac{\partial P(x,y,t)}{\partial t} = \frac{\partial^2 P(x,y,t)}{\partial x^2} + \frac{\partial^2 P(x,y,t)}{\partial y^2} - P(x,o,t)(1/r)P(o,y,t) \tag{28}$$

$$\left.\frac{\partial P(x,y,t)}{\partial x}\right|_{\substack{x = o \\ x = 1}} = 0 \tag{29}$$

We set the initial condition for filter covariance as

$$P(x,y,t_o) = 0.1 \cos \pi x \ \cos \pi y \tag{30}$$

and utilize the Fourier's seperation method in obtaining a numerical solution for eqn. (28) with boundary condition (29). For r = 1, Fig. 2 shows the filter covariance at the boundary as function of x. In Fig. 3 the covariance trajectory at various spatial locations is depicted.

APPENDIX

The Kalman-Bucy distributed filter for processing boundary measurements has been obtained. Some of the recent studies consider the measurement data obtained from the systems interior. The obvious generalization of the filtering problem is to process simultaneously the data obtained from the interior of the systems domain as well as that obtained from the boundary. This is presented next.

Let the system under consideration be governed by the following equations:

$$\frac{\partial X(x,t)}{\partial t} = S_x\, X(x,t) + W_d(x,t) \qquad x \in \Omega \tag{31}$$

$$L_\beta\, X(\beta,t) = W_b(\beta,t) \qquad \beta \in \partial\,\Omega \tag{32}$$

In eqns (31) and (32) $W_d(x,t)$ and $W_b(\beta,t)$ are zero-mean white noises respectively with covariance functions $Q_d(x,y,t)$ and $Q_b(\beta,\alpha,t)$ respectively.

The simultaneous measurements inside the interior and at the boundary of the system are governed by following relations:

$$Z(t) = \int_\Omega H_d(x,t)\, X(x,t)\, d\Omega_x + V_d(t) \qquad x \in \Omega \tag{33}$$

$$Z(\beta,t) = H_b(\beta,t)\, X(\beta,t) + V_b(\beta,t) \qquad \beta \in \partial\,\Omega \tag{34}$$

The noise vectors $V_d(t)$ and $V_b(\beta,t)$ are assumed to be zero-mean and white with covariance functions $R_d(t)$ and $R_b(\beta,\alpha,t)$ respectively. For the system defined by eqns. (31) through (34) a general type of Kalman-Bucy distributed filter can be defined.

$$\begin{aligned}\frac{\partial\hat{X}(x,t/t)}{\partial t} = {} & S_x\hat{X}(x,t/t) + \int_\Omega P(x,s,t)H_d'(s,t)R^{-1}(t)\,d\Omega_s \cdot \\ & \cdot\Big[Z(t) - \int_\Omega H_d(\sigma,t)\,\hat{X}(\sigma,t/t)\,d\Omega_\sigma\Big] \\ & + \int_{\partial\Omega}\int_{\partial\Omega} P(x,\eta,t)H_b'(\eta,t)R^{i}(\eta,\alpha,t)\cdot \\ & \cdot\,\big[Z(\alpha,t) - H_b(\alpha,t)\hat{X}(\alpha,t/t)\; d(\partial\Omega_\eta)\,d(\partial\Omega_\alpha)\big]\end{aligned} \tag{35}$$

with homogeneous boundary condition

$$L_\beta \hat{X}(\beta, t/t) = 0 \qquad (36)$$

and

$$\frac{\partial P(x,y,t)}{\partial t} = S_x P(x,y,t) + P(x,y,t)\, S_y'$$

$$-\int_\Omega \int_\Omega P(x,s,t) H_d'(s,t) R^{-1}(t) H_d(\sigma,t) P(\sigma,y,t)\, d\Omega_s\, d\Omega_\sigma$$

$$-\int_{\partial\Omega}\int_{\partial\Omega} P(x,\eta,t) H_b'(\eta,t) R^{i}(\eta,\alpha,t) H_b(\alpha,t) P(\alpha,y,t)\, d(\partial\Omega_\eta)\, d(\partial\Omega_\alpha)$$

$$+ \quad Q_d(x,y,t)$$

$$+\int_{\partial\Omega}\int_{\partial\Omega} B(x,t;\eta,t) Q_b(\eta,\alpha,t) B'(y,t;\alpha,t)\, d(\partial\Omega_\eta)\, d(\partial\Omega_\alpha) \qquad (37)$$

with boundary condition

$$L_\beta P(\beta,y,t) = \int_{\partial\Omega} Q_b(\beta,\alpha,t) B'(y,t;\alpha,t)\, d(\partial\Omega_\alpha) \qquad (38)$$

REFERENCES

[1] S.G. Tzafestas and J.M. Nightingale, "Optimal Filtering, Smoothing and Prediction in Linear Distributed Parameter Systems," Proc. IEE, Vol. 115, pp. 1207-1212, Aug. 1968.

[2] S R. Atre and S.S. Lamba, "Optimal Estimator for Linear Distributed Parameter Systems," Proc. IEEE Symposium on Adaptive Processes: Decision and Control, Austin, Texas, December, 1970.

[3] F.E. Thau, "On Optimum Filtering for a class of Linear Distributed Parameter Systems," Jour. Basic Engrg., Trans ASME, Vol. 91D, pp. 173-178, June 1969.

[4] P.K.C Wang, "Control of Distributed Parameter Systems," in "Advances in Control Systems," Vol. 1, Ed. C.T Leondes New York, Academic Press, 1964.

[5] M. Kim and H. Erzberger, "On the Design of Optimum Distributed Parameter System with Boundary Control Function," IEEE Trans on Auto Contr. Vol.AC-12, pp22-28, Feb. 1967.

[6] A.V. Balakrishnan and J L. Lions, "State Estimation for Infinite-Dimensional Systems," Jour. Comp. and Syst. Sc. Vol. 1, pp. 391-403, 1967.

[7] T. Kailath, "An innovation Approach to Least-Squares Estimation- Part I: Linear Filtering in Additive White Noise," IEEE Trans. on Auto. Contr., Vol AC-13, pp. 646-655, December 1968.

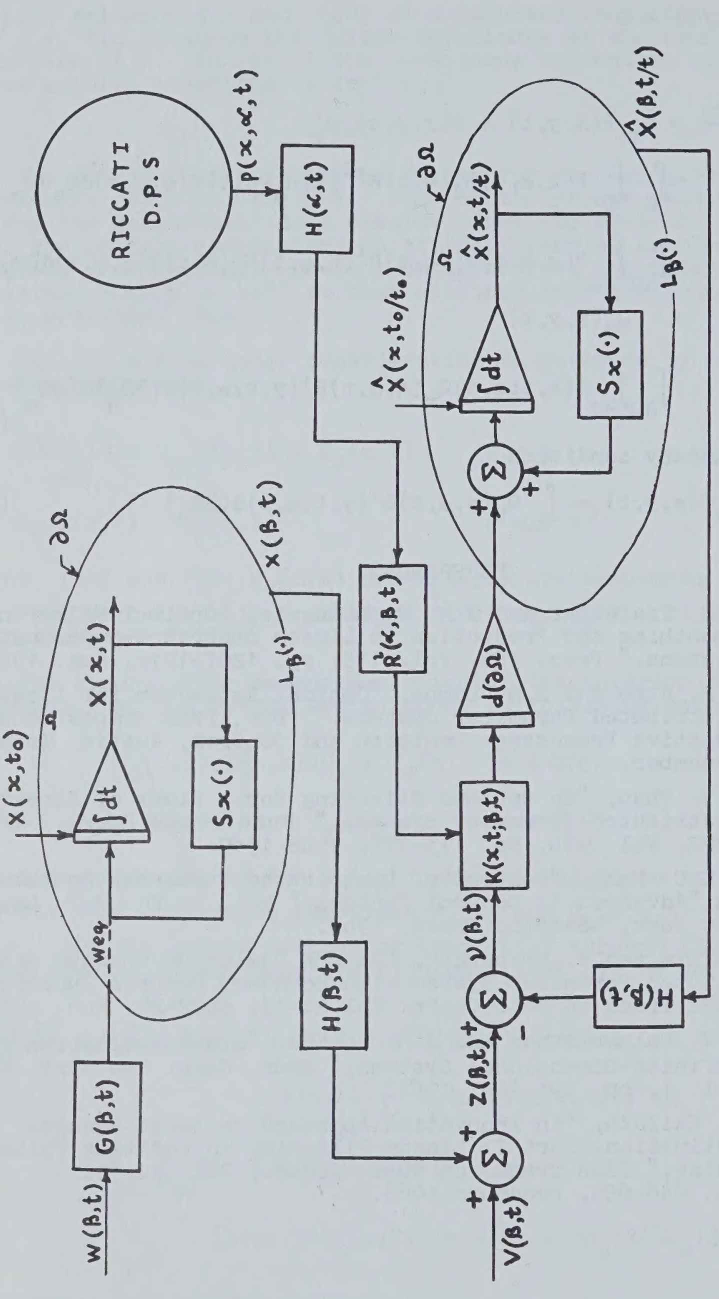

RICCATI
D.P.S
p(x,α,t)
H(α,t)
∂Ω
Ω
X̂(x,t/t)
X̂(x,t₀/t₀)
∫dt
Sx(·)
Σ
+
+
L_β(·)
X̂(β,t/t)
H(β,t)
R⁻¹(α,β,t)
d(∂Ω)
K(x,t;β,t)
ν(β,t)
Σ
+
-
Z(β,t)
Σ
+
+
V(β,t)
H(β,t)
W(β,t)
G(β,t)
We_q
X(x,t₀)
∫dt
Sx(·)
X(x,t)
Ω
∂Ω
L_β(·)
X(β,t)

Fig 1 Block Diagram of Boundary Filter

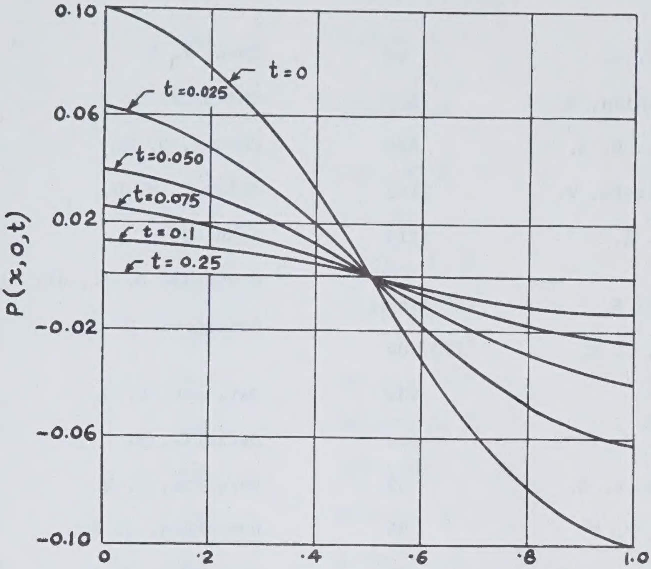


Fig. 2 Boundary Covariance as function of x

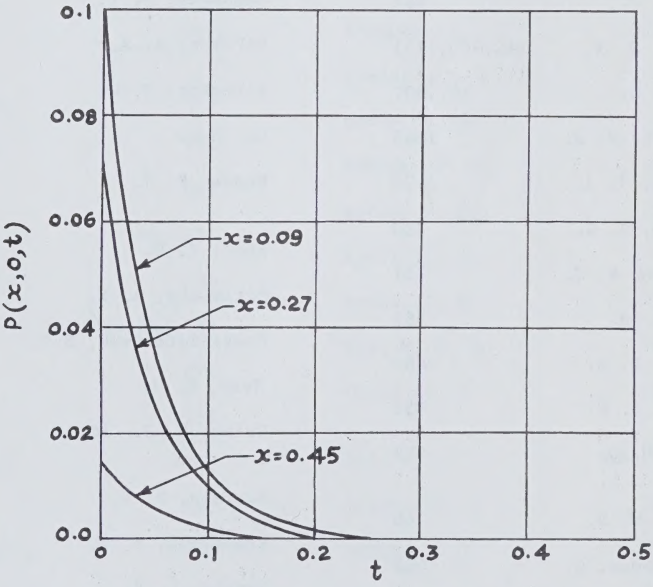


Fig. 3 Boundary Covariance as function of t

INDEX OF AUTHORS

Ahmed, N. U. 49
Aldermeshian, H. 845
Anderson, D. A. 696
Arandjelovic, V. 1193
Atre, S. R. 1213

Bajaj, K. S. 192,1180
Bandler, J. W. 170,404
Basar, T. 932
Bedford, J. S. 150
Bhagavan, B. K. 58
Bhatkar, V. P. 84
Bhattacharyya, B. B. 661
Bhushan, C. 661
Bickart, T. A. 142,465,1161
Bose, N. K. 157,901
Bouknight, W. J. 1045
Bourquin, J. J. 1170
Bredeson, J. G. 338
Brodersen, A. J. 131
Brown, W. A. 595
Burgess, D. A. 465
Burgiss, S. G. 330
Burton, H. O. 476

Ćalović, M. S. 69
Charalambous, C. 404
Chen, K. 827
Chen, Y. E. 93
Chien, R. T. 689
Chyung, D. H. 921
Ciletti, M. D. 911
Cohn, N. 719
Costello, D. J., Jr. 677
Cristiani, E. J. 864

Davisson, L. D. 477
DeClaris, N.
DeSantis, R. M. 434
Denenberg, J. N. 1085
Denenberg, S. A. 1010
Dessouky, M. I. 644
DiPerna, R. A. 474
Director, S. W. 131
Du, M-W. 122
Dunne, E. J. 644

Eger, T. E. 1076
Eltimsahy, A. H. 258
Evers-Euterneck, E.T. 624
Evey, R. J. 864
Ewing, D. J., Jr. 1203

Fairman, F. W. 424
Franaszek, P. A. 507
Franks, L. E. 222

Frost, P. A. 232
Fu, K. S. 951

Garcia, O. N. 709
Gardner, W. A. 222
Gault, J. W. 114
Gelenbe, E. 104
Gilbart, J. W. 604
Gibson, J. E. 538
Giorgi, C. G. 884
Goldman, R. E. 864
Gordon, H. 894
Greene, R. T. 709
Grigsby, L. L. 613
Gupta, M. M. 41
Gupta, N. N. 906

Hakimi, S. L. 455
Hanson, J. L. 942
Hao, H. T. 347
Happ, W. W. 652,1151
Harman, R. K. 424
Hartmann, C. R. P. 685,687,689
Hemami, H. 876
Hensley, R. D. 1170
Higgins, T. J. 1203
Ho, C. 140
Hoelscher, H. E. 827
Hofstetter, E. M. 789
Huang, T-C. 799
Hulina, P. T. 338
Hutton, M. F. 160

Isidori, A. 884

Jain, N. C. 634
Jan, Y. 140
Jensen, B. D. 212
Jermihov, N. 862

Kadota, T. T. 485
Kaiser, J. F. 809
Karmarkar, J. S. 585
Kashyap, R. L. 32
Kazda, L. F. 258
Kim, H. K. 675
Klapper, J. 1067
Klauder, L. T., Jr. 531
Koivo, H. N. 282
Kohler, W. H. 377
Kotob, S. M. 157
Krall, E. J. 309
Kreer, J. B. 525
Kuo, M. C. Y. 924
Kusic, G. L. 1193

Lal, M. 634,663
Lainiotis, D. G. 246
Lamba, S. S. 1213
Lee, C. 1140
Lee, H. E. 1124

Lee, S. C. 357
Leon, B. J. 799
Lind, R. O. 318
Lizewski, E. 604
Lugannani, R. 1094

Madzarević, V. 730
Mantey, P. E. 864
Marma, V. J. 514
Markettos, N. D. 170
McIntyre, D. E. 1039
McNamee, L. P. 652,1140
Meditch, J. S. 24
Meier, L. 397
Melvin, D. W. 142
Metwally, A. A. 576
Metze, G. 696
Mickle, M. H. 845
Miczo, A. 329
Milsum, J. 444
Mintz, M. 932
Mitra, S. K. 810
Mittra, R. 997
Monopoli, R. V. 604
Montgomery, R. C. 613
Mostafavi, M. 997
Moussa, H. 768
Murano, K. 1055

Nagy, G. 987
Nardizzi, L. R. 58
Neumann, H. J. 150
Newborn, M. M. 347
Nikiforuk, P. N. 41

Oka, K. 837
O'Neill, W. D. 855
Opačić, J. 369
Oppenheim, A. V. 789

Peach, L. C. 1085
Pereboom, E. G. 730
Phan, C. S. 675
Picon-Jimenez, J. L. 613
Porte, H. A. 1151
Porter, W. A. 434

Rajamani, V. S. 84
Randal, J. M. 1020
Resh, J. A. 1180
Ribak, R. 102
Ribbens-Pavella, M. 751
Root, W. L. 1
Rudolph, L. D. 329
Rugh, W. J. 866

Sabath, J. 730
Saeks, R. 180,290
Salmon, D. M. 272
Sameh, A. H. 1030
Sannuti, P. 182

Schatz, J. G. 1130
Schaubert, D. H. 997
Schwarzlander, H. 1112
Schwyn, P. J. 1161
Sea, R. G. 201
Shen, D. W. C. 1124
Shen, V. Y. 93
Sherwood, R. 810
Siegel, J. 789
Šiljak, D. D. 585,731
Singer, R. A. 201
Singh, H. 663
Singh, S. N. 866
Singhal, K. 1191
Sinha, N. K. 170
Sloate, H. M. 465
Slotnick, D. L. 1007
Smith, L. W. 299
Smith, R. H., Jr. 414
Snyder, D. L. 242
Sorenson, H. W. 11
Souza, C. R. 180
Staudhammer, J. 330
Stefanek, R. G. 547
Steiglitz, K. 377,778
Stern, T. E. 1130
Strohmaier, R. 444
Swamy, M. N. S. 661

Tacker, E. C. 82
Thomas, J. B. 1055
Tiberio, P. 104
Tretiak, O. J. 497
Turski, A. W. 538
Tzeng, K. K. 685,689

Uyeno, D. 444

Vandelinde, V. D. 414,1114
Veiga, G. M. 1067
Vertinsky, I. 444
Vlach, J. 1191
Vogt, W. G. 845

Wagner, L. 41
Wagner, W. S. 673
Wang, S. J. 82
Wayne, D. A. 131
Weiss, C. D. 122
Wilkie, D. F. 559
Will, H. 444
Willems, J. L. 742
Wozny, M. J. 390

Yao, K. 1104
Yau, S. S. 299,318
Yeh, R. T. 309
Yu, Y. N. 576,768

Zein, D. A. R. 455

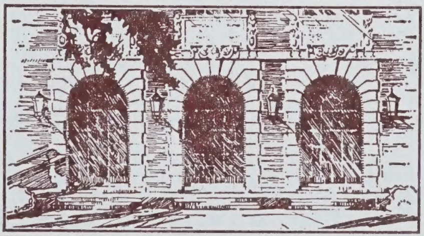
Zelenka, J. S. 975

Zobrist, G. W. 894

DEPARTMENT OF
COMPUTER SCIENCE
LIBRARY

LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

NOTICE: According to Sec. 19 (a) of the University *Statutes,* all books and other library materials acquired in any manner by the University belong to the University Library. When this item is no longer needed by the department, it should be returned to the Acquisition Department, University Library.



UNIVERSITY OF ILLINOIS-URBANA
621.31921AL54P C003
PROCEEDINGS.$URBANA ILL
9 1971
3 0112 006279662